BIENNIAL REPORT

2004 – 2005

Laboratory of Computer and Information Science

Neural Networks Research Centre

Helsinki University of Technology

P.O. Box 5400

FIN-02015 HUT, Finland

V. Könönen, K. Raivio, R. Vigário, and L. Koivisto, editors

Otaniemi, April 2006

ISSN 1795-5092
ISSN 1796-4121 (electronic version)

Yliopistopaino
Helsinki 2006

Contents

Preface

The Laboratory of Computer and Information Science (CIS, informaatiotekniikan laboratorio) is one of the research and teaching units of the Department of Computer Science and Engineering at Helsinki University of Technology. The laboratory has its roots in the Electronics Laboratory, established in 1965 by Professor Teuvo Kohonen. For more than 30 years, the research in the laboratory has concentrated on neurocomputing, especially associative memories, self-organization, and adaptive signal and image processing, as well as on their applications on pattern recognition. The laboratory has grown from its roots of one professor and a handful of students and researchers into a relatively large and well established unit of 5 professors and about 80 staff altogether.

The Neural Networks Research Centre (NNRC, neuroverkkojen tutkimusyksikkö) was established by Professor Kohonen in 1994 as a separate research unit with its own funding and own administrative position. It was selected as one of the first Finnish national Centers of Excellence in Research in 1995. The Academy of Finland extended its Center of Excellence status for the years 2000 to 2005 under the research proposal "New Information Processing Principles". This status has also implied financial resources from the Academy, Tekes, HUT, and Nokia Co., which are gratefully acknowledged.

The present biennial report includes the description of the final two years of the Neural Networks Research Centre. The activities of the centre did not end in 2005, though; they were inherited by a new and larger research unit, whose name "Adaptive Informatics Research Centre" reflects the changing emphasis of the research conducted here. This unit, too, was selected by the Academy of Finland as a national Center of Excellence for the years 2006 - 2011.

The Neural Networks Research Centre has operated within the Laboratory of Computer and Information Science, coordinating the major part of its research activities. It is not possible to separate the personnels of these two units, as the teaching staff of the CIS also participated in some research project of the NNRC. Professor Erkki Oja was the director of NNRC, with Professor Olli Simula the vice-director, and Professor Juha Karhunen participated in its research projects. In addition, 18 post-doctoral researchers, 34 graduate students, and a number of undergraduate students were working in the NNRC projects in 2005.

Professor Heikki Mannila joined the CIS laboratory in 1999. He is partner and vice-director of the **From Data to Knowledge research unit (FDK, Datasta tietoon - tutkimusyksikkö)**, a joint effort between Helsinki University of Technology and the University of Helsinki. Also this research group was selected as a national Center of Excellence from the beginning of 2002. Although the Neural Networks Research Centre and the From Data to Knowledge research unit were financially separate and stemmed from different research traditions, there has been an overlap in the research directions and projects between these two Centers of Excellence. This overlap has produced fruitful joint research which is expected to increase in the future.

The present report covers the activities during the years 2004 and 2005. Basically, the report is divided in two parts. In the first part, the research of the NNRC is reviewed. In the second part, those projects of the FDK research unit are reviewed, that pertain to the research activities in the CIS laboratory. The main reason for this separation is that the present booklet also serves as the official report of the NNRC to its sponsors, and it is important to clearly distinguish exactly what work has been done under those finances.

The earlier achievements and developments of the NNRC have been thoroughly explained in the triennial reports 1994 - 1996 and 1997 - 1999 for the first period of CoE status, as well as the biennial reports 2000 - 2001 and 2002 - 2003 for the first four years of the second period. The web pages of the laboratory, `http://www.cis.hut.fi/` also contain up-to-date texts.

To briefly list the main numerical achievements of the period 2004 - 2005, the laboratory produced 9 D.Sc. (Eng.) degrees, 4 Lic.Tech. degrees, and 42 M.Sc. (Eng.) degrees. The number of scientific publications appearing during the period was 244, of which 51 were journal papers. It can be also seen that the impact of our research is clearly increasing, measured by the citation numbers to our previously published papers and books, as well as the number of users of our public domain software packages.

A large number of talks, some of them plenary and invited, were given by our staff in the major conferences in our research field. We had several foreign visitors participating in our research, and our own researchers made visits to universities and research institutes abroad. The research staff were active in international organizations, editorial boards of journals, and conference committees. Also, some prices and honours, both national and international, were granted to members of our staff.

One of the highlights of the two-year period was the 40 year anniversary of the Laboratory, held on May 3, 2005. It was in spring 1965 that Professor Teuvo Kohonen was appointed to his chair.

Erkki Oja	*Olli Simula*	*Heikki Mannila*
Professor	Professor	Academy Professor
Director,	Director,	Vice Director,
Neural Networks	Laboratory of Computer	From Data to Knowledge
Research Centre	and Information Science	Research Unit

Personnel

Employees during 2004 – 2005

Professors

Erkki Oja, D.Sc. (Tech.), Academy Professor. Director, Neural Networks Research Centre
Olli Simula, D.Sc. (Tech.). Director, Laboratory of Computer and Information Science
Teuvo Kohonen, D.Sc. (Tech.), Emeritus Professor, Academician
Juha Karhunen, D.Sc. (Tech.), part-time from Feb. 2005
Heikki Mannila, PhD, Academy Professor. Vice Director, From Data to Knowledge research unit
Samuel Kaski, D.Sc. (Tech.), from July 2005 (up to July '05 at University of Helsinki)
Timo Honkela, PhD, until Aug. 2005, Acting Professor (E. Oja's chair)
Jaakko Hollmén, D.Sc. (Tech.), Acting Professor (H. Mannila's chair)

Post-doc researchers

Ella Bingham, D.Sc. (Tech.), until Aug. 2005
Robert Gwadera, PhD, from Sept. 2005
Johan Himberg, D.Sc. (Tech.), until Sept. 2005
Antti Honkela, D.Sc. (Tech.)
Timo Honkela, PhD
Jukka Iivarinen, D.Sc. (Tech.)
Mika Inki, D.Sc. (Tech.)
Sirkka-Liisa Joutsiniemi, D.Med.Sc., Feb. - March 2004
Markus Koskela, D.Sc. (Tech.), visiting abroad from Nov. 2005
Mikko Kurimo, D.Sc. (Tech.)
Ville Könönen, D.Sc. (Tech.)
Jorma Laaksonen, D.Sc. (Tech.)
Krista Lagus, D.Sc. (Tech.)
Sampsa Laine, D.Sc. (Tech.), until end of 2004
Amaury Lendasse, PhD
Janne Nikkilä, D.Sc. (Tech.)
Petteri Pajunen, D.Sc. (Tech.)
Kalle Palomäki, D.Sc. (Tech.), from April 2005
Jaakko Peltonen, D.Sc. (Tech.)
Kai Puolamäki, PhD
Kimmo Raivio, D.Sc. (Tech.)
Janne Sinkkonen, D.Sc. (Tech.), until Aug. 2004
Miki Sirola, D.Sc. (Tech.), laboratory engineer
Panu Somervuo, D.Sc. (Tech.), until end of 2004

Jaakko Särelä, D.Sc. (Tech.)
Harri Valpola, D.Sc. (Tech.), until Nov. 2004
Ricardo Vigário, D.Sc. (Tech.)

Post-graduate researchers

Matti Aksela
Mathias Creutz
Ramunas Girdziusas
Teemu Hirsimäki
Alexander Ilin
Heli Hiisilä
Arto Klami
Mikko Koivisto, until July 2004
Leo Lahti, absent Aug. 2004 - Aug. 2005
Pasi Lehtimäki
Merja Oja
Jussi Pakkanen
Jukka Parviainen
Anne Patrikainen
Astrid Pietilä, from March 2005
Janne Pylkkönen
Matti Pöllä
Tapani Raiko
Karthikesh Raju, until end of 2005
Ann Russell, May - Dec. 2005
Salla Ruosaari
Jarkko Salojärvi
Eerika Savia
Jouni Seppänen
Timo Similä
Vesa Siivola
Harri Sulkava
Nikolai Tatti
Jarkko Tikka
Antti Ukkonen
Jarkko Venna
Sampo Viiperi, until July 2004
Ville Viitaniemi
Jaakko Väyrynen
Zhirong Yang
Jarkko Ylipaavalniemi
Zhijian Yuan

Under-graduate researchers (full-time or part-time)

Phani Sudheer Bhutadi, until June 2005
Simo Broman, until March 2005
Jin Hao, until end of 2005
Markus Harva
Mikko Heikelä, until April 2004

Hannes Heikinheimo
Kevin Hynnä
Yongnan Ji
Jari-Pekka Juhala, until end of 2004
Oskar Kohonen, until July 2005
Lauri Kovanen
Jakke Kulovesi
Mikko Korpela
Kalle Korpiaho, until Sept. 2004
Mikaela Kumlander
Golan Lampi, abroad Oct. 2004 - July 2005
Elia Liitiäinen
Tiina Lindh-Knuutila
Leo Lundqvist, until Sept. 2004
Teppo Marin, until Nov. 2005
Mikko Multanen
Hannes Muurinen
Antti Puurula
Antti Rasinen
Rami Rautkorpi
Ulpu Remes
Nima Reyhani
Kosti Rytkönen, until Sept. 2005
Petri Saarikko, until end of 2005
Antti Savolainen, until Sept. 2004
Jan-Hendrik Schleimer
Santeri Seppä
Mats Sjöberg
Antti Sorjamaa
Janne Toivola
Matti Tornio
Ville Turunen
Matti Varjokallio
Sami Virpioja
Paul Wagner
Yujie Ye, until end of 2005
Tomas Östman, until end of 2004

Support staff

Leila Koivisto, department secretary
Sakari Laitinen, maintenance assistant, until June 2005
Tapio Leipälä, maintenance assistant
Tarja Pihamaa, laboratory secretary
Markku Ranta, B.Eng., works engineer
Petteri Räisänen, maintenance assistant
Jaakko Salomaa
Mika Kongas

Awards and activities

Prizes and scientific honours received by researchers of the unit

Prof. Erkki Oja:

- Pierre Devijver Award, International Association for Pattern Recognition (IAPR), USA. 2004.
- Unsupervised learning ICA pioneer award, SPIE defence and security symposium, USA. 2004.

Dr. Mikko Kurimo:

- International Short Visit Fellowship Award, The Royal Society, U.K. 2004.

Dr. Jorma Laaksonen, M.Sc. Ville Viitaniemi, Dr. Markus Koskela:

- Best CIS Paper Award at the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05), Finland. 2005.

M.Sc. Matti Pöllä and Professor Timo Honkela:

- Best Paper Award (soft computing, computational intelligence, fuzzy systems, neural networks, learning) at the Seventh International Conference on Computing Anticipatory Systems, Belgium. 2005.

Important international positions of trust held by researchers of the unit

Academy Professor Erkki Oja:

- Editorial Board Member:
  Neural Computation, USA
  International Journal of Pattern Recognition and Artificial Intelligence, Singapore
  Natural Computing–An International Journal, The Netherlands.

- President of European Neural Network Society (ENNS), The Netherlands.

- Evaluator:
  EU 6th Framework Program, Belgium, 2004
  Deutsche Forschungsgemeinshaft DFG, Germany, 2004
  Norges Forskningsråd, Norway, 2004.

- Governing Board Member of International Neural Network Society (INNS), USA.

- Opponent at the doctoral dissertation of Jacob Verbeek, University of Amsterdam, The Netherlands, 2004.

- Opponent at the doctoral dissertation of Robert Jenssen, University of Tromsö, Norway, 2005.

- Opponent at the doctoral dissertation of Hamed Hamid Muhammed, Uppsala University, Sweden, 2005.

- Honorary Chairman of Engineering of Intelligent Systems (EIS) conference, Madeira, Portugal, Feb. 29–Mar. 2, 2004.

- Invited talk “Applications of Independent Component Analysis.” International Conference on Neural Information Processing (ICONIP’04), Calcutta, India, Nov. 22-25, 2004.

- Plenary talk “Blind Source Separation: Neural Net Principles and Applications.” SPIE Defense and Security Symposium, Orlando, USA, Apr. 12-16, 2004.

- Plenary talk “Finding Clusters and Components by Unsupervised Learning.” IAPR International Workshop on Statistical Pattern Recognition (SPR2004), Lisbon, Portugal, Aug. 18-20, 2004.

- Plenary talk “Patterns, Clusters, and Components - What Data is Made of?” International Joint Conference on Neural Networks (IJCNN’04), Budapest, Hungary, Jul. 26-29, 2004.

- Plenary talk “Independent Component Analysis: Algorithms and Applications.” International Joint Conference on Robust Statistics (ICORS), Jyväskylä, Finland, Jun. 14, 2005.

- Plenary talk “Finding Hidden Factors in Large Spatiotemporal Data Sets.” 2005 International Conference on Neural Networks and Brain, Beijing, China, Oct. 13 - 15, 2005.

- Plenary talk “Finding Hidden Factors in Large Spatiotemporal Data Sets.” CERCIA Workshop on Computational Intelligence, Birmingham, UK, Nov. 15–16, 2005.

- Plenary talk "Finding Hidden Factors in Large Spatiotemporal Data Sets." DEST Workshop on Machine Learning for Applications in Sensor Networks, Int. Conf. on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), Melbourne, Australia, Dec. 5–8, 2005.

- Tutorial talk "Independent Component Analysis: Theory and Applications." 14th Scandinavian Conference on Image Analysis (SCIA), Joensuu, Finland, Jun. 19, 2005.

- Conference co-chair of International Conference on Neural Information Processing (ICONIP), Calcutta, India, Nov. 22–25, 2004.

- Session Chairman and Program Committee Member:
  5th Workshop on Self-Organizing Maps (WSOM 05), Paris, France, Sep. 5–8, 2005
  International Conference on Artificial Neural Networks (ICANN 2005), Warsaw, Poland, Sep. 11–15, 2005.

- Program Committee Member:
  International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, Jul. 25–29, 2004
  European Symposium on Artificial Neural Networks (ESANN), Bruges, Belgium, Apr. 28–30, 2004
  International Conference on Pattern Recognition (ICPR), Cambridge, U.K., Aug. 23–26, 2004
  IEEE Workshop on Machine Learning for Signal Processing (MLSP), Sao Luis, Brazil, Sep. 29–Oct. 1, 2004
  5th International Conference on Independent Component Analysis and Blind Signal Separation (ICA), Granada, Spain, Sep. 22–24, 2004.

- Co-Editor
  Artificial Neural Networks: Biological Inspirations, Lecture Notes in Computer Science 3696, Berlin, Germany, 2005
  Artificial Neural Networks: Formal Models and Their Applications, Lecture Notes in Computer Science 3697, Berlin, Germany, 2005.

Professor Olli Simula:

- Executive Board Member of European Neural Network Society (ENNS), The Netherlands.

- Chairman of the Finnish IEEE–Computer Chapter.

- Scientific Council Member of Institute Eurecom, France.

- Program Committee Member:
  International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, Jul. 25–29, 2004.
  5th Workshop on Self-Organizing Maps (WSOM 05), Paris, France, Sep. 5–8, 2005
  International Conference on Artificial Neural Networks (ICANN 2005), Warsaw, Poland, Sep. 11–15, 2005.

Professor Juha Karhunen:

- Editorial Board Member of Neurocomputing, The Netherlands.

- Program Committee Member:
  5th International Conference on Independent Component Analysis and Blind Signal Separation (ICA), Granada, Spain, Sep. 22–24, 2004
  13th European Symposium on Artificial Neural Networks (ESANN'2005), Bruges, Belgium, Apr. 27–29, 2005.

Professor Heikki Mannila:

- Editor-in-Chief, Data Mining and Knowledge Discovery, USA, 2004.
- Area Editor, IEEE Transactions on Knowledge and Data Engineering, USA, 2004.
- Co-Editor, *Constraint–based mining and inductive databases*, Lecture Notes in Computer Science 3848, Berlin, Germany, 2005.
- Member of Steering Committee, Data Mining and Knowledge Discovery, USA, 2005.
- Associate Editor, ACM Transactions on Database Systems, USA, 2005.
- Member of Editorial Board, ACM Transactions on Knowledge Discovery in Data, USA, 2005.
- Member of Technical Advisory Board, Verity Inc., USA, 2004.
- Member of ESFRI Physical Sciences and Engineering Roadmap Working Group, Luxembourg, 2005.
- Chairman of ESFRI Expert Group on Computation and Data Treatment, Luxembourg, 2005.
- Member of PKDD Steering Committee, USA, 2004.
- Program Committee Member, Tenth SIGKDD Conference on Data Mining and Knowledge Discovery (KDD'04), Seattle, Washington, USA, 2004.
- Member of ACM SIGKDD Curriculum Committee, USA, 2005.

Professor Jaakko Hollmén:

- Member of the Management Board in the Knowledge Discovery Network of Excellence (KDNet) supported by EU Project No. IST-2001-33086.
- Program Committee Member:
  16th European Conference on Artificial Intelligence (ECAI 2004), Valencia, Spain, Aug. 22–27 2004
  7th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2005), Copenhagen, Denmark, Aug. 22–26, 2005.

Professor Timo Honkela:

- Representative of Finland of International Federation on Information Processing (IFIP), TC12 (Artificial Intelligence), Austria, 2004.
- Chairman of International Federation on Information Processing (IFIP), WG12.1 (Knowledge Representation and Reasoning), Austria.

- EU Commission, Expert in FP7 preparations in the area of cognitive systems, Belgium, 2005.
- Opponent at the doctoral dissertation of Apostolos A. Georgakis, Umeå University, Sweden, 2004.
- Invited talk "Translation Within and Between Languages." European Conference on Computing and Philosophy (E-CAP'05), Västerås, Sweden, Jun. 4, 2005.

Academician Teuvo Kohonen:

- Plenary talk "Pointwise Organizing Projections." 5th Workshop On Self-Organizing Maps (WSOM'05), Paris, France, Sep. 5-8, 2005.

Professor Samuel Kaski:

- Editorial Board Member:
  International Journal of Neural Systems, Singapore
  Intelligent Data Analysis, The Netherlands
  Cognitive Neurodynamics, Germany, 2005.
- Member of the Machine Learning for Signal Processing Technical Committee of the IEEE Signal Processing Society, USA, 2004.
- Referee of Fund for Scientific Research-Flanders, Belgium, 2004.
- Referee of Engineering and Physical Sciences Research Council, U.K., 2004.
- Referee of Netherlands Organization for Scientific Research, The Netherlands, 2004.
- Opponent at the doctoral dissertation of Catherine Aaron, Universite de Paris 1, France, 2005.
- Opponent at the doctoral dissertation of Shaun Mahony, National University of Ireland, Ireland, 2005.
- Plenary talk "From Learning Metrics Towards Dependency Exploration." 5th Workshop On Self-Organizing Maps (WSOM'05), Paris, France, Sep. 5-8, 2005.
- Invited talk "Proactive Information Retrieval by Monitoring Eye Movements." BrainIT, The Second Int. Conf. on Brain-Inspired Information Technology, Kyushu Institute of Technology, Kitakuyshu, Japan, Oct. 7-9, 2005.
- Chairman and Program Committee Member in NIPS 2005 Workshop on Machine Learning for Implicit Feedback and User Modeling, Whistler, Canada, Dec. 10, 2005.
- Session Chairman in:
  IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005), Philadelphia, USA, Mar. 18–23, 2005
  5th Workshop on Self-Organizing Maps (WSOM 05), Paris, France, Sep. 5–8, 2005.
- Program Committee Member:
  New Trends in Intelligent Information Processing and Web Mining (IIPWM'05), Zakopane, Poland, Jun. 13–15, 2005
  2005 Atlantic Web Intelligence Conference (AWIC'05), Lodz, Poland, Jun. 6–9, 2005

IEEE International Workshop on Machine Learning for Signal Processing (MLSP'05), Mystic, Connecticut, USA, Sep. 28–30, 2005
The 2005 IEEE/WIC International Conference on Web Intelligence (WI 2005), Compiègne, France, Sep. 19–22, 2005
International Conference on Natural Computation (ICNC'05), Hunan, China, Aug. 27–29, 2005
4th International Workshop on Web Semantics (WebS 2005), Copenhagen, Denmark, Aug. 22–26, 2005
Workshop at UM'05, Machine Learning for User Modeling: Challenges, Edingburgh, U.K., Jul. 23–29, 2005
5th Workshop on Self-Organizing Maps (WSOM 05), Paris, France, Sep. 5–8, 2005
International Symposium on Intelligent Data Engineering and Automated Learning (IDEAL'05), Brisbane, Australia, Jul. 6–8, 2005
European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'05), Porto, Portugal, Oct. 3–7, 2005.

Dr. Jukka Iivarinen:

- Program Committee Member:
  International Conference on Artificial Intelligence and Soft Computing, Marbella, Spain, Sep. 1–3, 2004
  International Conference on Computational Intelligence, Calgary, Canada, Jul. 4–6, 2005.

Dr. Mikko Kurimo:

- Session Chairman in XII. European Signal Processing Conference EUSIPCO, Vienna, Austria, Sep. 6–10, 2004.

Dr. Ville Könönen:

- Session Chairman and Program Committee Member of the Workshop 9 in the 9th European Conference on Machine Learning (ECML 2005), Porto, Portugal, Oct. 7–9, 2005.

Dr. Amaury Lendasse:

- Invited talk "Time Series Benchmarking Competition: The CATS Benchmark." International Joint Conference on Neural Networks (IJCNN'04), Budapest, Hungary, Jul. 26–29, 2004.

Dr. Miki Sirola:

- Session Chairman and Program Committee Member:
  IEEE International Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS'2005), Sofia, Bulgaria, Sep. 5–7, 2005.

- Program Committee Member:
  International Conference on Modelling and Simulation, Marina del Ray, USA, Mar. 1–3, 2004

International Conference on Applied Simulation and Modelling, Rhodes, Greece, Jun. 28–30, 2004
International Conference on Intelligent Systems and Control, Honolulu, Hawaii, Aug. 23–25, 2004
International Conference on Neural Information Processing (ICONIP), Calcutta, India, Nov. 22–25, 2004
International Conference on Modeling, Identification and Control, Grindelwald, Switzerland, Feb. 22-25, 2005
International Conference on Modelling and Simulation, Cancun, Mexico, May. 18–20, 2005
International Conference on Modelling, Identification and Control, Innsbruck, Austria, Feb. 16–18, 2005
International Conference on Applied Simulation and Modelling, Benalmádena, Spain, Jun. 15–17, 2005
International Conference on Computational Intelligence, Calgary, Canada, Jul. 4–6, 2005
International Conference on Intelligent Systems and Control, Cambridge, USA, Oct. 31–Nov. 2, 2005.

Dr. Ricardo Vigário:

- Plenary talk "Search for Independence in Biomedical Systems." 2nd Int. Conf. on Advances in Medical Signal and Information Processing (MEDSIP'04), Malta, Sep. 5-8, 2004.

- Invited talk "Single Trial Denoising Source Separation of Event-Related Fields" co-authored by Jaakko Särelä. Tandem Workshop on "Advanced Methods of Electrophysiological Signal Analysis" & "Symbolic Grounding? Dynamical Systems Approaches in Language", Potsdam, Germany, Mar. 2005.

- Program Committee Member:
  5th International Conference on Independent Component Analysis and Blind Signal Separation (ICA), Granada, Spain, Sep. 22–24, 2004
  2nd International Conference on Computational Intelligence in Medicine and Healthcare (CIMED'05), Portugal, Jul. 2005.

M.Sc. Mika Sulkava:

- Invited talk: "Laboratory Quality Affects the Ability to Detect Temporal Changes in Foliar Element Concentrations." 9th ICP Forest Expert Panel on Foliar Analysis, Dublin, Ireland, Jun. 20, 2005. Part of International Co-operative Programme on Assessment and Monitoring of Air Pollution Effects on Forests of UN/ECE (ICP Forests) in cooperation with the EU.

Important domestic positions of trust held by researchers of the unit

Academy Professor Erkki Oja:

- Evaluator in filling the academic chairs of professors:
  computer science, University of Joensuu, 2004
  mathematics, Tampere University of Technology, 2005.
- Session Chairman in International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05), Espoo, Jun. 15–17, 2005.

Professor Olli Simula:

- Evaluator in filling the academic chairs of professors:
  computer engineering, University of Jyväskylä, 2004
  electronics, Tampere University of Technology, 2004.
- Chairman of the Organizing Committee in International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05), Espoo, Jun. 15–17, 2005.
- Panel Moderator in 2005 IEEE Mid-Summer Workshop on Soft-Computing in Industrial Applications (SMCia 05), Espoo, Jun. 28–30, 2005.

Professor Timo Honkela:

- Vice Chairman, Finnish Cognitive Linguistics Association, 2004.
- Chairman of the Program Committee in International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05), Espoo, Jun. 15–17, 2005.
- Editor Board Member, journal Puhe ja kieli (Speech and Language), 2005.
- Opponent at the doctoral dissertation of Wanjiku Ng'ang'a, University of Helsinki, 2005.

Professor Samuel Kaski:

- Session Chairman and Program Committee Member in Symposium on Knowledge Representation in Bioinformatics (KRBIO'05), Espoo, Jun. 15, 2005.
- Co-Editor, Proceedings of Symposium on Knowledge Representation in Bioinformatics (KRBIO 05), 2005.
- Opponent at the doctoral dissertation of Harri Lähdesmäki, Tampere University of Technology, 2005.
- Opponent at the doctoral dissertation of Mantao Xu, University of Joensuu, 2005.

Professor Jaakko Hollmén:

- Opponent at the doctoral dissertation of Tomas Eklund, Åbo Akademi, 2004.
- Opponent at the doctoral dissertation of Kari Vasko, University of Helsinki, 2004.

- Co-Editor, Proceedings of Symposium on Knowledge Representation in Bioinformatics (KRBIO 05), 2005.
- Program Committee Member in Symposium on Knowledge Representation in Bioinformatics (KRBIO 05), 2005.

Dr. Jorma Laaksonen:

- Finnish Artificial Intelligence Society. Chairman of the dictionary committee.

Dr. Jukka Iivarinen:

- Pattern Recognition Society of Finland. Vice chairman.

Dr. Ricardo Vigário:

- Session Chairman in International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05), Espoo, Jun. 15–17, 2005.

Dr. Ann Russell:

- Session Chairman in International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05), Espoo, Jun. 15–17, 2005.

Petri Saarikko:

- Program Committee Member in International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05), Espoo, Jun. 15–17, 2005.

M.Sc. Matti Pöllä:

- Program Committee Member in International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05), Espoo, Jun. 15–17, 2005.

Dr. Krista Lagus:

- Invited talk "Miten hermoverkkomallit selittävät kielen oppimista." Puheen ja kielen tutkimuksen yhdistyksen päivät, March 17–18, 2005, Finland.
- Session Chairman and Program Committee Member in International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05), Espoo, Jun. 15–17, 2005.

Dr. Ville Könönen:

- Session Chairman and Program Committee Member in International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05), Espoo, Jun. 15–17, 2005.

M.Sc. Tiina Lindh–Knuutila:

- Secretary of the Program Committee in International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05), Espoo, Jun. 15–17, 2005.

Dr. Mikko Kurimo:

- Session Chairman in International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05), Espoo, Jun. 15–17, 2005.

Dr. Kimmo Raivio:

- Opponent at the doctoral dissertation of Eero Wallenius, University of Jyväskylä, 2005.

M.Sc. Tapani Raiko:

- Chairman of Finnish Artificial Intelligence Society, 2005.

Research visits abroad by researchers of the unit; 2 weeks or more

- Dr. Ella Bingham, University of Birmingham, Sep. 2004–Dec. 2004.
- M.Sc. Mathias Creutz, International Computer Science Institute, USA, Nov. 2005–.
- M.Sc. Markus Harva, University of Birmingham, Jun. 2004–Aug. 2004; Mar. 2005–Apr. 2005.
- Dr. Samuel Kaski, Université Paris 1, France, June 2004.
- Dr. Markus Koskela, Dublin City University, Ireland, Nov. 2005–.
- Dr. Mikko Kurimo, University of Edinburgh, U.K., Jun. 2004–Jul. 2004.
- M.Sc. Merja Oja, Uppsala University, Feb. 2004–Mar. 2004.
- Lic.Sc. Anne Patrikainen, University of Washington, USA, Feb. 2004–Oct. 2004; Jan. 2005–Jun. 2005.
- M.Sc. Salla Ruosaari, University Medical Center Hamburg–Eppendorf, Germany, May, 2005 (2 wks).
- Dr. Harri Valpola, University of Zurich, Switzerland, Sep. 2003–Oct. 2004.

Research visits by foreign researchers to the unit; 2 weeks or more

- Milenko Adamovic, University of Banjaluka, Bosnia-Herzegovina, Jun. 2005–Jul. 2005.
- Mariana Almeida, Technical University of Lisbon, Portugal, Feb. 2005–Jul. 2005.
- M.Sc. Ebru Arisoy, Bogazici University, Turkey, Jul. 2005–Aug. 2005.
- Dr. Patrik Bas, Institut National Polytechnique de Grenoble, France, Sep. 2004–.
- B.Sc. Phani Sudheer Bhutadi, National Institute of Technology, India, Jan. 2004–May 2005.
- Dr. Sergey Borisov, Moscow State University, Russia, Jun. 2004–Nov. 2005.
- M.Sc. Basilio Calderone, Scuola Normale Superiore, Italy, Feb. 2005–Jul. 2005.
- M.Sc. Francesco Corona, Universita degli Studi di Cagliari, Italy, Jan. 2005–Jun. 2005; Nov. 2005–Dec. 2005.
- Prof. César Fernández, Miguel Hernandez University, Spain, Aug. 2004–Nov. 2004.
- Dr. Robert Gwadera, Purdue University, USA, Sep. 2005–.
- Milan Jurik, Czech Technical University, Czech Republic, Sep. 2004–Dec. 2004.
- Prof. Sami Khuri, San José State University, USA, Apr. 2004–Jun. 2004.

- M.Sc. Jan Kriz, Technical University Liberec, Czech Republic, Sep. 2005–.
- Dr. Amaury Lendasse, Université Catholique de Louvain, Belgium, Jan. 2004–.
- Prof. Nobuo Matsuda, Oshima College of Maritime Technology, Japan, Nov. 2005–.
- M.Sc. Nasser Mourad, South Valley University, Egypt, Oct. 2003–Jan. 2004.
- Prof. Bryan Pellom, University of Colorado at Boulder, USA, Aug. 2004–Dec. 2004.
- B.Sc. Nima Reyhani, University of Tehran, Iran, Oct. 2004–.
- Dr. Ann Russel, University of Toronto, Canada, May 2005–Dec. 2005.
- B.Sc. Jan-Hendrik Schleimer, University of Tübingen, Germany, Aug. 2003–.

Other activities

Professor Olli Simula:

- interview on subject ”ICT education at Finnish universities” for the journal IT-Viikko/IT News, Nov. 2005.

Professor Samuel Kaski:

- interview for Aamulehti (one of the main Finnish newspapers) on data mining, May 13, 2004.
- article in Verkkouutiset (web newspaper) about genomics data mining, www.verkkouutiset.fi, May 28, 2004, ”Geenitietoa louhitaan oppivilla menetelmillä.”

Dr. Mikko Kurimo:

- interview on ”How does a machine understand speech?” for the science series program Prisma Studio, YLE TV1, March 31, 2004.

Dr. Jorma Laaksonen:

- appearance in popular science TV program on biometric person identification (in Finnish), Prisma Studio, YLE TV1, Feb. 2005.

Dr. Ella Bingham, Dr. Ricardo Vigário, M.Sc. Jarkko Ylipaavalniemi:

- interview on subject ”independent component analysis” for the journal Tekniikka & Talous, May 26, 2005, p. 44.

Dr. Jaakko Särelä, M.Sc. Alexander Ilin, Dr. Harri Valpola (HUT Computational Eng. Lab.):

- interview on ”denoising source separation and its applications to climatogy” was published in several national newspapers and radio programmes, June 2005.

Courses

Courses given by the Laboratory of Computer and Information Science.

Spring Semester 2004

Code	Course	Lecturer	Course Assistant
T-61.140	Signal Processing Systems	O. Simula	J. Parviainen, A. Sorjamaa, V. Viitaniemi
T-61.152	Seminar on Computer and Information Science	E. Bingham	
T-61.233	Computer Vision	J. Laaksonen	J. Iivarinen
T-61.256	Learning Models and Methods	P. Pajunen	E. Bingham
T-61.261	Principles of Neural Computing	K. Raivio, T. Honkela	J. Venna, M. Aksela
T-61.281	Statistical Natural Language Processing	T. Honkela K. Lagus	V. Siivola
T-61.182	Special Course II: *Information Theory and Machine Learning*	J. Karhunen	A. Honkela
T-61.183	Special Course III: *Bioinformatics*	S. Khuri, J. Hollmén	H. Hiisilä
T-122.102	Special Course VI: *Co-occurence Methods in Analysis of Discrete Data*	K. Puolamäki, J. Hollmén	A. Klami

Fall Semester 2004

Code	Course	Lecturer	Course Assistant
T-61.123	Computer Architecture	S. Haltsonen	L. Eloranta, M. Mäkinen, A. Sorjamaa
T-61.124	Special Project in Computer Architecture	S. Haltsonen	
T-61.231	Principles of Pattern Recognition	T. Honkela K. Raivio M. Koskela	M. Koskela M. Aksela
T-61.238	Statistical Signal Modelling	P. Pajunen	V. Viitaniemi
T-61.246	Digital Signal Processing and Filtering	O. Simula	J. Parviainen, K. Korpiaho, A. Savolainen
T-61.247	Digital Image Processing	J. Laaksonen	J. Iivarinen
T-61.263	Advanced Course in Neural Computing	J. Karhunen	J. Peltonen
T-61.271	Information Visualisation	K. Puolamäki	T. Raiko
T-61.181	Special Course I: *Biomedical Signal Processing*	R. Vigário, J. Särelä	
T-61.184	Special Course IV: *Speech Recognition and Language Modeling - From Theory to Practice*	B. Pellom	
T-61.186	Special Course in Language Techonology: *The Evolution of Language and Cognition: Modeling Perspectives*	K. Lagus, K. Kanto, T. Honkela	
T-122.101	Special Course V: *Analysis of Time-series and Sequences*	A. Lendasse, J. Hollmén	A. Sorjamaa

Spring Semester 2005

Code	Course	Lecturer	Course Assistant
T-61.140	Signal Processing Systems	O. Simula	J. Parviainen, K. Korpiaho A. Savolainen
T-61.152	Seminar on Computer and Information Science: *Causality*	E. Bingham	
T-61.233	Computer Vision	J. Laaksonen	J. Iivarinen
T-61.256	Learning Models and Methods	P. Pajunen	V. Viitaniemi
T-61.261	Principles of Neural Computing	K. Raivio	J. Venna M. Aksela
T-61.281	Statistical Natural Language Processing	K. Lagus, T. Honkela	S. Virpioja
T-61.182	Special Course II: *Biomedical Image Analysis*	R. Vigário, J. Särelä	
T-61.183	Special Course II: *Multimodal Systems*	T. Honkela, M. Kurimo, J. Laaksonen, K. Lagus, K. Puolamäki	
T-122.102	Special Course VI: *Regularization and Sparse Approximations*	J. Hollmén, A. Lendasse	J. Tikka

Fall Semester 2005

Code	Course	Lecturer	Course Assistant
T-61.123	Computer Architecture	S. Haltsonen	L. Eloranta, M. Mäkinen, A. Sorjamaa
T-61.124	Special Project in Computer Architecture	S. Haltsonen	
T-61.2010	From Data to Knowledge	H. Mannila E. Oja	J. Parviainen U. Remes
T-61.3040	Statistical Signal Modelling	P. Pajunen	V. Viitaniemi
T-61.5030	Advanced Course in Neural Computing	J. Karhunen	J. Peltonen
T-61.5050	High-Throughput Bioinformatics	S. Kaski P. Auvinen	J. Nikkilä
T-61.5060	Algorithmic methods of data mining	H. Mannila	K. Puolamäki
T-61.5080	Signal Processing in Neuroinformatics	R. Vigário J. Särelä	
T-61.5100	Digital Image Processing	J. Laaksonen	J. Iivarinen
T-61.6050	Special Course V: *Neural Networks for Modelling and Control of Dynamic Systems*	J. Hollmén A. Lendasse	J. Tikka
T-61.6080	Special Course in Bioinformatics: *Analysis of proteomic and metabolic data*	S. Kaski M. Oresic	J. Nikkilä
T-61.6090	Special Course in Language Technology: *Statistical Machine Translation*	T. Honkela	T. Lindh-Knuutila J. Väyrynen

Doctoral dissertations

Data exploration with self-organizing maps in environmental informatics and bioinformatics

Mikko Kolehmainen

Dissertation for the degree of Doctor of Science in Technology on 27 February 2004.

External examiners:
Enso Ikonen (University of Oulu)
Erkki Pesonen (University of Kuopio)
Opponent:
Jussi Parkkinen (University of Joensuu)

Abstract:

The aim of this thesis was to evaluate the usability of self-organizing maps and some other methods of computational intelligence in analysing and modelling problems of environmental informatics and bioinformatics. The concepts of environmental informatics, bioinformatics, computational intelligence and data mining are first defined. There follows an introduction to the data processing chain of knowledge discovery and the methods used in this thesis, namely linear regression, self-organizing maps (SOM), Sammon's mapping, U-matrix representation, fuzzy logic, c-means and fuzzy c-means clustering, multi-layer perceptron (MLP), and regularization and Bayesian techniques. The challenges posed by environmental processes and bioprocesses are then identified, including missing data problems, complex lagged dependencies among variables, non-linear chaotic dynamics, ill-defined inverse problems, and large search space in optimization tasks.

The works included in this thesis are then evaluated and discussed. The results show that the combination of SOM and Sammon's mapping has great potential in data exploration, and can be used to reveal important features of the measurement techniques (e.g. separability of compounds), reveal new information about already studied phenomena, speed up research work, act as a hypothesis generator for traditional research, and supply clear and intuitive visualization of the environmental phenomenon studied. The results of regression studies show, as expected, that the MLP network yields better estimates in predicting future values of airborne pollutant concentration of NO_2 compared with SOM based regression or the Least Squares approach using periodic components. Additionally, the use of local MLP models is shown to be slightly better for estimating future values of episodes compared with one MLP model only. However, it can be concluded in general that the architectural issues tested are not able to solve solely model performance problems.

Finally, recommendations for future work are laid out. Firstly, the data exploration solution should be enhanced with methods from signal processing to enable the handling of measurements with different time scale and lagged multivariate time-series. The main suggestion, however, is to create an integrated environment for testing different hybrid schemes of computational intelligence for better time-series forecasting in environmental informatics and bioinformatics.

Exploratory source separation in biomedical systems

Jaakko Särelä

Dissertation for the degree of Doctor of Science in Technology on 29 October 2004.

External examiners:
Te-Won Lee (University of California at San Diego)
Ole Jensen (Radboud University Nijmegen)
Opponent:
Lars Kai Hansen (Technical University of Denmark)

Abstract:
Contemporary science produces vast amounts of data. The analysis of this data is in a central role for all empirical sciences as well as humanities and arts using quantitative methods. One central role of an information scientist is to provide this research with sophisticated, computationally tractable data analysis tools.

When the information scientist confronts a new target field of research producing data for her to analyse, she has two options: She may make some specific hypotheses, or guesses, on the contents of the data, and test these using statistical analysis. On the other hand, she may use general purpose statistical models to get a better insight into the data before making detailed hypotheses.

Latent variable models present a case of such general models. In particular, such latent variable models are discussed where the measured data is generated by some hidden sources through some mapping. The task of *source separation* is to recover the sources. Additionally, one may be interested in the details of the generation process itself.

We argue that when little is known of the target field, *independent component analysis* (ICA) serves as a valuable tool to solve a problem called *blind source separation* (BSS). BSS means solving a source separation problem with no, or at least very little, prior information. In case more is known of the target field, it is natural to incorporate the knowledge in the separation process. Hence, we also introduce methods for this incorporation. Finally, we suggest a general framework of *denoising source separation* (DSS) that can serve as a basis for algorithms ranging from almost blind approach to highly specialised and problem-tuned source separation algoritms. We show that certain ICA methods can be constructed in the DSS framework. This leads to new, more robust algorithms.

It is natural to use the accumulated knowledge from applying BSS in a target field to devise more detailed source separation algorithms. We call this process *exploratory source separation* (ESS). We show that DSS serves as a practical and flexible framework to perform ESS, too.

Biomedical systems, the nervous system, heart, etc., constitute arguably the most complex systems that human beings have ever studied. Furthermore, the contemporary physics and technology have made it possible to study these systems while they operate in near-natural conditions. The usage of these sophisticated instruments has resulted in a massive explosion of available data. In this thesis, we apply the developed source separation algorithms in the analysis of the human brain, using mainly magnetoencephalograms (MEG). The methods are directly usable for electroencephalograms (EEG) and with small adjustments for other imaging modalities, such as (functional) magnetic resonance imaging (fMRI), too.

From insights to innovations: data mining, visualization, and user interfaces

Johan Himberg

Dissertation for the degree of Doctor of Science in Technology on 5 November 2004.

External examiners:
Sami Khuri (San José State University)
Olli Silvén (University of Oulu)
Opponent:
Juha Röning (University of Oulu)

Abstract:
This thesis is about data mining (DM) and visualization methods for gaining insight into multidimensional data. Novel, exploratory data analysis tools and adaptive user interfaces are developed by tailoring and combining existing DM and visualization methods in order to advance in different applications.

The thesis presents new visual datamining (VDM)methods that are also implemented in software toolboxes and applied to industrial and biomedical signals: First, we propose a method that has been applied to investigating industrial process data. The self-organizing map (SOM) is combined with scatterplots using the traditional color linking or interactive brushing. The original contribution is to apply color linked or brushed scatterplots and the SOM to visually survey local dependencies between a pair of attributes in different parts of the SOM. Clusters can be visualized on a SOM with different colors, and we also present how a color coding can be automatically obtained by using a proximity preserving projection of the SOM model vectors. Second, we present a new method for an (interactive) visualization of cluster structures in a SOM. By using a contraction model, the regular grid of a SOM visualization is smoothly changed toward a presentation that shows better the proximities in the data space. Third, we propose a novel VDM method for investigating the reliability of estimates resulting from a stochastic independent component analysis (ICA) algorithm. The method can be extended also to other problems of similar kind. As a benchmarking task, we rank independent components estimated on a biomedical data set recorded from the brain and gain a reasonable result.

We also utilize DMand visualization for mobile-awareness and personalization. We explore how to infer information about the usage context from features that are derived from sensory signals. The signals originate from a mobile phone with on-board sensors for ambient physical conditions. In previous studies, the signals are transformed into descriptive (fuzzy or binary) context features. In this thesis, we present how the features can be transformed into higher-level patterns, contexts, by rather simple statistical methods: we propose and test using minimum-variance cost time series segmentation, ICA, and principal component analysis (PCA) for this purpose. Both time-series segmentation and PCA revealed meaningful contexts from the features in a visual data exploration.

We also present a novel type of adaptive soft keyboard where the aim is to obtain an

ergonomically better, more comfortable keyboard. The method starts from some conventional keypad layout, but it gradually shifts the keys into new positions according to the user's grasp and typing pattern.

Related to the applications, we present two algorithms that can be used in a general context: First, we describe a binary mixing model for independent binary sources. The model resembles the ordinary ICA model, but the summation is replaced by the Boolean operator OR and the multiplication by AND. We propose a new, heuristic method for estimating the binary mixing matrix and analyze its performance experimentally. The method works for signals that are sparse enough. We also discuss differences on the results when using different objective functions in the FastICA estimation algorithm. Second, we propose "global iterative replacement" (GIR), a novel, greedy variant of a merge-split segmentation method. Its performance compares favorably to that of the traditional top-down binary split segmentation algorithm.

Data exploration with learning metrics

Jaakko Peltonen

Dissertation for the degree of Doctor of Science in Technology on 17 November 2004.

External examiners:
Hannu Toivonen (University of Helsinki)
Kari Torkkola (Motorola Labs)
Opponent:
John Shawe-Taylor (University of Southampton)

Abstract:

A crucial problem in exploratory analysis of data is that it is difficult for computational methods to focus on interesting aspects of data. Traditional methods of unsupervised learning cannot differentiate between interesting and noninteresting variation, and hence may model, visualize, or cluster parts of data that are not interesting to the analyst. This wastes the computational power of the methods and may mislead the analyst.

In this thesis, a principle called "learning metrics" is used to develop visualization and clustering methods that automatically focus on the interesting aspects, based on auxiliary labels supplied with the data samples. The principle yields non-Euclidean (Riemannian) metrics that are data-driven, widely applicable, versatile, invariant to many transformations, and in part invariant to noise.

Learning metric methods are introduced for five tasks: nonlinear visualization by Self-Organizing Maps and Multidimensional Scaling, linear projection, and clustering of discrete data and multinomial distributions. The resulting methods either explicitly estimate distances in the Riemannian metric, or optimize a tailored cost function which is implicitly related to such a metric. The methods have rigorous theoretical relationships to information geometry and probabilistic modeling, and are empirically shown to yield good practical results in exploratory and information retrieval tasks.

Linear space–time modulation in multiple–antenna channels

Ari Hottinen

Dissertation for the degree of Doctor of Science in Technology on 25 November 2004.

External examiners:
Jyrki Joutsensalo (University of Jyväskylä)
Tapani Ristaniemi (Tampere University of Technology)
Opponent:
David Gespert (Institut Eurécom)

Abstract:

This thesis develops linear space–time modulation techniques for (multi-antenna) multi-input multi-output (MIMO) and multiple-input single-output (MISO) wireless channels. Transmission methods tailored for such channels have recently emerged in a number of current and upcoming standards, in particular in 3G and ‘’beyond 3G” wireless systems. Here, these transmission concepts are approached primarily from a signal processing perspective.

The introduction part of the thesis describes the transmit diversity concepts included in the WCDMA and cdma2000 standards or standard discussions, as well as promising new transmission methods for MIMO and MISO channels, crucial for future high data-rate systems. A number of techniques developed herein have been adopted in the 3G standards, or are currently being proposed for such standards, with the target of improving data rates, signal quality, capacity or system flexibility.

The thesis adopts a model involving matrix-valued modulation alphabets, with different dimensions usually defined over *space* and *time*. The symbol matrix is formed as a linear combination of symbols, and the space-dimension is realized by using multiple transmit and receive antennas. Many of the transceiver concepts and modulation methods developed herein provide both spatial multiplexing gain and diversity gain. For example, full-diversity full-rate schemes are proposed where the symbol rate equals the number of transmit antennas. The modulation methods are developed for open-loop transmission. Moreover, the thesis proposes related closed-loop transmission methods, where space–time modulation is combined either with automatic retransmission or multiuser scheduling.

Multiagent reinforcement learning: asymmetric and symmetric approaches

Ville Könönen

Dissertation for the degree of Doctor of Science in Technology on 3 December 2004.

External examiners:
Petri Koistinen (University of Helsinki)
Kary Främling (Helsinki University of Technology)
Opponent:
Ann Nowé (Vrije Universiteit Brussel)

Abstract:

Modern computing systems are distributed, large, and heterogeneous. Computers, other information processing devices and humans are very tightly connected with each other and therefore it would be preferable to handle these entities more as agents than stand-alone systems. One of the goals of artificial intelligence is to understand interactions between entities, whether they are artificial or natural, and to suggest how to make good decisions while taking other decision makers into account. In this thesis, these interactions between intelligent and rational agents are modeled with Markov games and the emphasis is on adaptation and learning in multiagent systems.

Markov games are a general mathematical tool for modeling interactions between multiple agents. The model is very general, for example common board games are special instances of Markov games, and particularly interesting because it forms an intersection of two distinct research disciplines: machine learning and game theory. Markov games extend Markov decision processes, a well-known tool for modeling single-agent problems, to multiagent domains. On the other hand, Markov games can be seen as a dynamic extension to strategic form games, which are standard models in traditional game theory. From the computer science perspective, Markov games provide a flexible and efficient way to describe different social interactions between intelligent agents.

This thesis studies different aspects of learning in Markov games. From the machine learning perspective, the focus is on a very general learning model, i.e. reinforcement learning, in which the goal is to maximize the long-time performance of the learning agent. The thesis introduces an asymmetric learning model that is computationally efficient in multiagent systems and enables the construction of different agent hierarchies. In multiagent reinforcement learning systems based on Markov games, the space and computational requirements grow very quickly with the number of learning agents and the size of the problem instance. Therefore, it is necessary to use function approximators, such as neural networks, to model agents in many real-world applications. In this thesis, various numeric learning methods are proposed for multiagent learning problems.

The proposed methods are tested with small but non-trivial example problems from different research areas including artificial robot navigation, simplified soccer game, and automated pricing models for intelligent agents. The thesis also contains an extensive literature survey on multiagent reinforcement learning and various methods based on Markov games.

Extensions of independent component analysis for natural image data

Mika Inki

Dissertation for the degree of Doctor of Science in Technology on 10 December 2004.

External examiners:
Michael Lewicki (Carnegie Mellon University)
Heikki Hyötyniemi (Helsinki University of Technology)
Opponent:
Gustavo Deco (Universitat Pompeu Fabra)

Abstract:

An understanding of the statistical properties of natural images is useful for any kind of processing to be performed on them. Natural image statistics are, however, in many ways as complex as the world which they depict. Fortunately, the dominant low-level statistics of images are sufficient for many different image processing goals. A lot of research has been devoted to second order statistics of natural images over the years.

Independent component analysis is a statistical tool for analyzing higher than second order statistics of data sets. It attempts to describe the observed data as a linear combination of independent, latent sources. Despite its simplicity, it has provided valuable insights of many types of natural data. With natural image data, it gives a sparse basis useful for efficient description of the data. Connections between this description and early mammalian visual processing have been noticed.

The main focus of this work is to extend the known results of applying independent component analysis on natural images. We explore different imaging techniques, develop algorithms for overcomplete cases, and study the dependencies between the components by using a model that finds a topographic ordering for the components as well as by conditioning the statistics of a component on the activity of another. An overview is provided of the associated problem field, and it is discussed how these relatively small results may eventually be a part of a more complete solution to the problem of vision.

Advances in variational Bayesian nonlinear blind source separation

Antti Honkela

Dissertation for the degree of Doctor of Science in Technology on 13 May 2005.

External examiners:
Fabian Theis (University of Regensburg)
Aki Vehtari (Helsinki University of Technology)
Opponent:
Tom Heskes (Radboud University Nijmegen)

Abstract:

Linear data analysis methods such as factor analysis (FA), independent component analysis (ICA) and blind source separation (BSS) as well as state-space models such as the Kalman filter model are used in a wide range of applications. In many of these, linearity is just a convenient approximation while the underlying effect is nonlinear. It would therefore be more appropriate to use nonlinear methods.

In this work, nonlinear generalisations of FA and ICA/BSS are presented. The methods are based on a generative model, with a multilayer perceptron (MLP) network to model the nonlinearity from the latent variables to the observations. The model is estimated using variational Bayesian learning. The variational Bayesian method is well-suited for the nonlinear data analysis problems. The approach is also theoretically interesting, as essentially the same method is used in several different fields and can be derived from several different starting points, including statistical physics, information theory, Bayesian statistics, and information geometry. These complementary views can provide benefits for interpretation of the operation of the learning method and its results.

Much of the work presented in this thesis consists of improvements that make the nonlinear factor analysis and blind source separation methods faster and more stable, while being applicable to other learning problems as well. The improvements include methods to accelerate convergence of alternating optimisation algorithms such as the EM algorithm and an improved approximation of the moments of a nonlinear transform of a multivariate probability distribution. These improvements can be easily applied to other models besides FA and ICA/BSS, such as nonlinear state-space models. A specialised version of the nonlinear factor analysis method for post-nonlinear mixtures is presented as well.

Exploratory cluster analysis of genomic high-throughput data sets and their dependencies

Janne Nikkilä

Dissertation for the degree of Doctor of Science in Technology on 1 December 2005.

External examiners:
Olli Yli-Harja (Tampere University of Technology)
Matej Oresic (Technical Research Centre of Finland)
Opponent:
Alvis Brazma (European Bioinformatics Institute)

Abstract:

This thesis studies exploratory cluster analysis of genomic high-throughput data sets and their interdependencies. In modern biology, new high-throughput measurements generate numerical data simultaneously from thousands of molecules in the cell. This enables a new perspective to biology, which is called systems biology. The discipline developing methods for the analysis of the systems biology data is called bioinformatics. The work in this thesis contributes mainly to bioinformatics, but the approaches presented are general purpose machine learning methods and can be applied in many problem areas.

A main problem in analyzing genomic high-throughput data is that the potentially useful new findings are hidden in a huge data mass. They need to be extracted and visualized to the analyst as overviews.

This thesis introduces new exploratory cluster analysis methods for extracting and visualizing findings of high-throughput data. Three kinds of methods are presented to solve progressively better-focused problems. First, visualizations and clusterings using the self-organizing map are applied to genomic data sets. Second, the recently developed methods for improving the visualization and clustering of a data set with auxiliary data are applied. Third, new methods for exploring the dependency between data sets are developed and applied. The new methods are based on maximizing the Bayes factor between the model of independence and the model of dependence for finite data.

The methods outperform their alternatives in numerical comparisons. In applications they proved capable of confirming known biological findings, which validates the methods, and also generated new hypotheses. The applications included exploration of yeast gene expression data, yeast gene expression data in a new metric learned with auxiliary data, the regulation of yeast gene expression by transcription factors, and the dependencies between human and mouse gene expression.

Theses

Licentiate of Science in Technology

2004

Lehtola, Aarno
Grammar formalism for controlled language machine translation: Augmented lexical entries

Reunanen, Juha
Overfitting pitfalls in feature selection

Vilkama, Esa
Integration of learning systems and prior knowledge for manufacturing process

2005

Patrikainen, Anne
Methods for comparing subspace clusterings

Master of Science in Technology

2004

Gävert, Hugo
Bankruptcy prediction and cluster analysis of small and medium-sized enterprises based on financial statements

Harva, Markus
Hierarchical variance models of image sequences

Heinonen, Mikko
Video transcoding using digital signal processor

Kinnunen, Tero
Design and implementation of a licensing system for remote patient monitorin software

Kontio, Juho
Neuroevolution based artificial bandwidth expansion of telephone band speech

Korkiakoski, Visa
Parallel simulation of a coronagraph for extremely large telescopes

Laaksonen, Miika
Akustinen tiedonsiirtoprotokolla ja sen luotettavuuden parantaminen (Acoustic data transimission protocol and its reliability)

Lundén Petteri
Analysis of the wideband CDMA on the 2.6 GHz extension band

Paavola, Risto
Paperikoneen poikkiradan kutistumamittaus digitaalisen kuvankäsittelyn avulla (Paper mill cross section shrinkage measurement using digital image processing)

Pylkkönen, Janne
Phone duration modeling techniques in continuous speech recognition

Rasinen, Lasse
Aikakauslehtien myyntipistekohtainen valikoimanoptimointi (Magazine sales outlet assortment optimization)

Similä, Timo
The impact of research and development on growth in Finnish manufacturing firms

Simola, Juhani
Protein identification by tandem mass spectrometry and sequence database search

Talvinen, Tea
Konfiguroitavan kliinisen kemian automaatiojärjestelmän toiminnan monitoroinnin kehittäminen (Developing monitoring of the functioning of a configurable clinical chemistry automation system)

Tatti, Nikolaj
Dissimilarity measures between binary data sets

Teräs, Arto
Database for ground based snowfall observation

Tikka, Jarkko
Learning linear dependency trees from multivariate data

Ukkonen, Antti
Data mining techniques for discovering partial orders

Verta, Heikki
Trace based off-line analysis of component-based systems

2005

Broman, Simo
Combination methods for language models in speech recognition

Hao, Jin
Input selection using mutual information - Applications to time series prediction

Hanhijärvi, Sami
Methods of active learning with model selection

Heikinheimo, Hannes
Inferring taxonomic hierarchies from 0-1 data

Hinneri, Samuli
Visual attention detection of Gaussian profile 2D structures in medical imaging

Ji, Yongnan
Least squares support vector machines for time series prediction

Kohonen, Oskar
Generering av naturligt språk från emergenta representationer (Natural language generation using emergent representations)

Knuuttila, Juha
Analysis of selected in silico haplotyping methods in presence of missing genotype data and linkage disequilibrium between genetic SNP markers

Kokkola, Antti
Optimization of noise suppression algorithm for digital signal processors

Kärkkäinen, Anssi
Menetelmiä radiotaajuisen mittausaineiston analyysiin ja visualisointiin (Methods for analysis and visualization of radio frequency measurement data)

Lagutin, Dmitrij
An interactive tool for structuring user data

Liitiäinen, Elia
Stochastic nonlinear filtering in continuous time

Lindh-Knuutila, Tiina
Simulating the emergence of a shared conceptual system in a multi-agent environment

Marin, Teppo
A data analysis tool set to support the data mining process

Mäkelä, Riikka
Methods for bilateral haemodynamic activation studies of newborn infants

Nyrkkö, Seppo
RDF-mallien käyttö ontologiapohjaisessa dialogin hallinnassa (Utilisation of RDF-models in ontology-based dialogue management)

Phani Sudheer, Bhutadi
Interference cancellation in CDMA systems – Various approaches

Pöllä, Matti
Modeling anticipatory behavior with self-organizing neural networks

Rautkorpi, Rami
Shape features in the classification and retrieval of surface defect images

Sorjamaa, Antti
Strategies for the long-term prediction of time series using local models

Turunen, Ville
Spoken document retrieval in Finnish based morpheme-like subword units

Vesalainen, Heikki
Automatic extraction of protein-protein interactions from medical papers

Virpioja, Sami
New methods for statistical natural language modeling

Väyrynen, Jaakko
Learning linguistic features from natural text data by independent component analysis

Ylipaavalniemi, Jarkko
Variability of independent components in functional magnetic resonance imaging

I Neural Networks Research Centre Research Projects

Chapter 1

Introduction

The core area of research in the Neural Networks Research Centre (1994 - 2005) has been neurocomputing. We have traditions dating back to late 1960's in some areas like associative memories, learning algorithms, and self-organization, as well as related methods in pattern recognition.

By the 2000's, the field of neurocomputing has experienced considerable changes compared to its pioneering days. Most of the early artificial neural network models, now classics in the field, were strongly motivated by insights from neurobiology. Even today, there is a strong research effort in biologically motivated neural models and computational neuroscience. Some work along those lines has been conducted in our laboratory, too.

However, aside from this, another part of the field has developed into purely computational science and engineering that has very few, if any, connections to biology. These two directions, that of neuroscience and that of computational science, have largely diverged and found their own research societies. Prompted by some urgent new problems in information sciences, the computational methods have merged with other related fields like advanced statistics, pattern recognition, signal and data analysis, machine learning, and artificial intelligence, to a new field sometimes termed *statistical machine learning.* This is the major research area in our research today. Also the range of application fields has grown from pattern recognition and control to cover many new disciplines such as bioinformatics, neuroinformatics, and multimedia data analysis. To better emphasize this paradigm change, the name of our research unit was changed to "Adaptive Informatics Research Centre" as of January 1, 2006.

Pattern analysis and statistical machine learning are the central tools for structuring raw information, which needs to be filtered and restructured before it becomes usable. Techniques that can quickly analyze complex patterns and adapt to new data will be indispensable for maintaining a competitive edge in information-intensive applications. The basic scientific problem is to build empirical models of complex systems, based on natural or real-world data. The goal is to understand better the underlying phenomena, structures, and patterns buried in the large or huge data sets. Real-world data means e.g. images, sounds, speech, or measurements, contrary to symbolic data like text. However, today the statistical machine learning methods are migrating into the analysis of symbolic data, too, such as large text collections, Web pages, or genomic sequence data, exhibiting real-world complexities and ambiguities. If the datasets are large enough, even the symbolic data can in many cases be analyzed with statistical methods, complementing the conventional string processing or grammar-based algorithms.

Natural data has properties such as nonlinearity, nongaussianity, and complex interactions that have not been taken into account in classical multivariate statistics. Therefore, such models must be based on new information processing principles. In our approach, the intrinsic latent features or components of the observations, and their mutual interrelations, are learned from the data using automated machine learning methods. In this way, we build data-driven statistical models of the complex systems and structures that underlie natural or real-world data. The goal is to understand better the underlying phenomena, structures, and patterns buried in the large or huge data sets, in order to make the information usable.

In the Neural Networks Research Centre and its successor Adaptive Informatics Research Centre, we develop such models, study their theoretical properties, and apply them to problems in signal, image, and data analysis. All the work is based on the core expertise stemming from our own scientific inventions. The most classic of these are the Self-Organizing Map (SOM), introduced by Prof. Kohonen in early 1980's, and new learning algorithms for Principal/Independent Component Analysis which have been intensively studied in the 1990's. Both have been thoroughly covered in a large number of articles and books and have been extensively cited. Our present research largely builds on these methods.

Our focus is to create and maintain research groups with internationally recognized status. Figure 1.1 is a concise description of our internal project organization during 2004 - 2005. The Research Unit consisted of 3 major research groups, each having a number of projects. Typically, these project groups consist of senior researchers, graduate students, and undergraduate students. The number of doctor-level researchers in the NNRC (Dec. 2005) was 20, and of full-time graduate student researchers 34. This kind of organizational chart necessarily gives a very strict and frozen view of the research activities. The topics of the projects are heavily overlapping and there is a continuous exchange of ideas and sometimes researchers between the projects. In the following Chapters, all of these projects are covered in detail.

Additional information including demos etc. is available from our Web pages, `www.cis.hut.fi/research`.

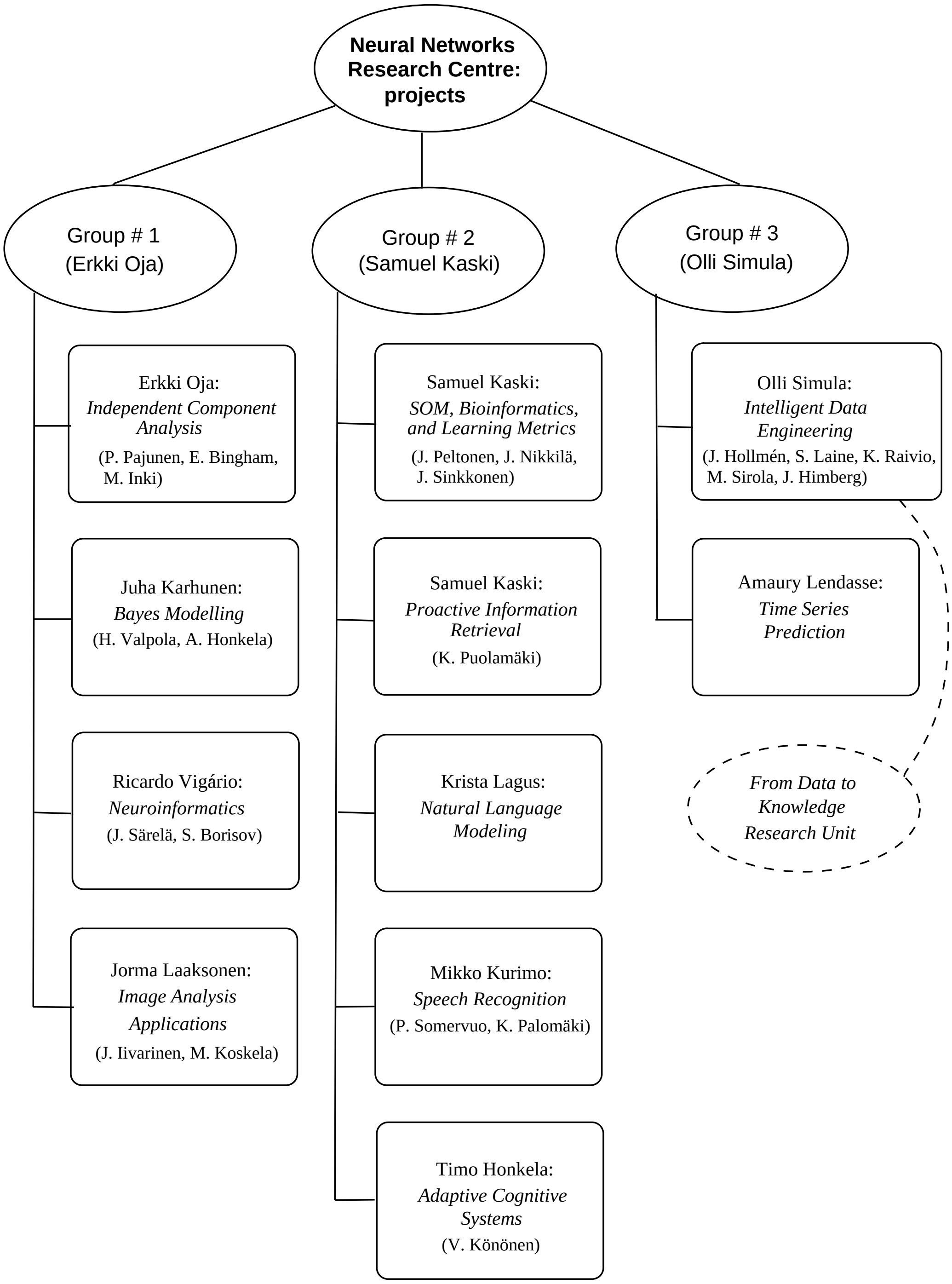

Neural Networks Research Centre: projects
Group # 1 (Erkki Oja)
Group # 2 (Samuel Kaski)
Group # 3 (Olli Simula)
Erkki Oja: Independent Component Analysis (P. Pajunen, E. Bingham, M. Inki)
Juha Karhunen: Bayes Modelling (H. Valpola, A. Honkela)
Ricardo Vigário: Neuroinformatics (J. Särelä, S. Borisov)
Jorma Laaksonen: Image Analysis Applications (J. Iivarinen, M. Koskela)
Samuel Kaski: SOM, Bioinformatics, and Learning Metrics (J. Peltonen, J. Nikkilä, J. Sinkkonen)
Samuel Kaski: Proactive Information Retrieval (K. Puolamäki)
Krista Lagus: Natural Language Modeling
Mikko Kurimo: Speech Recognition (P. Somervuo, K. Palomäki)
Timo Honkela: Adaptive Cognitive Systems (V. Könönen)
Olli Simula: Intelligent Data Engineering (J. Hollmén, S. Laine, K. Raivio, M. Sirola, J. Himberg)
Amaury Lendasse: Time Series Prediction
From Data to Knowledge Research Unit

Figure 1.1: The Neural Networks Research Centre in 2004 - 2005 consisted of three major groups, each having a number of smaller project groups. The leader of each group and the research topic are marked within each box, as well as the names of the post-doctoral researchers within each project. The dotted line indicates co-operation with the other Center of Excellence in the CIS laboratory.

Chapter 2

Independent component analysis and blind source separation

Erkki Oja, Juha Karhunen, Harri Valpola, Jaakko Särelä, Mika Inki, Antti Honkela, Alexander Ilin, Karthikesh Raju, Tapani Ristaniemi, Ella Bingham

2.1 Introduction

What is Independent Component Analysis and Blind Source Separation? Independent Component Analysis (ICA) is a computational technique for revealing hidden factors that underlie sets of measurements or signals. ICA assumes a statistical model whereby the observed multivariate data, typically given as a large database of samples, are assumed to be linear or nonlinear mixtures of some unknown latent variables. The mixing coefficients are also unknown. The latent variables are nongaussian and mutually independent, and they are called the independent components of the observed data. By ICA, these independent components, also called sources or factors, can be found. Thus ICA can be seen as an extension to Principal Component Analysis and Factor Analysis. ICA is a much richer technique, however, capable of finding the sources when these classical methods fail completely.

In many cases, the measurements are given as a set of parallel signals or time series. Typical examples are mixtures of simultaneous sounds or human voices that have been picked up by several microphones, brain signal measurements from multiple EEG sensors, several radio signals arriving at a portable phone, or multiple parallel time series obtained from some industrial process. The term blind source separation is used to characterize this problem. Also other criteria than independence can be used for finding the sources.

Our contributions in ICA research. In our ICA research group, the research stems from some early work on on-line PCA, nonlinear PCA, and separation, that we were involved with in the 80's and early 90's. Since mid-90's, our ICA group grew considerably. This earlier work has been reported in the previous Triennial and Biennial reports of our laboratory from 1994 to 2003. A notable achievement from that period was the textbook "Independent Component Analysis" (Wiley, May 2001) by A. Hyvärinen, J. Karhunen, and E. Oja. It has been very well received in the research community; according to the latest publisher's report, over 3900 copies have been sold by August, 2005. The book has been extensively cited in the ICA literature and seems to have evolved into the standard text on the subject worldwide. In 2005, the Japanese translation of the book appeared.

Another tangible contribution has been the public domain FastICA software package (`http://www.cis.hut.fi/projects/ica/fastica/`). This is one of the few most popular ICA algorithms used by the practitioners and a standard benchmark in algorithmic comparisons in ICA literature.

In the reporting period 2004 - 2005, ICA/BSS research stayed as a core project in the laboratory. It was extended to several new directions. This Chapter starts by introducing some theoretical advances on FastICA undertaken during the reporting period. Then, several extensions and applications of ICA and BSS are covered, namely nonlinear ICA and BSS, the Denoising Source Separation (DSS) algorithm, its applications to climate data analysis and telecommunications signals, ICA for image representations, and a latent variable method for analyzing binary data.

2.2 Finite sample behaviour of the FastICA algorithm

In ICA, a set of original source signals are retrieved from their mixtures based on the assumption of their mutual statistical independence. The simplest case for ICA is the instantaneous linear noiseless mixing model. In this case, the mixing process can be expressed as

$$\mathbf{X} = \mathbf{A}\mathbf{S}, \tag{2.1}$$

where $\mathbf{X}$ is an $d \times N$ data matrix. Its rows are the observed mixed signals, thus d is the number of mixed signals and N is their length or the number of samples in each signal. Similarly, the unknown $d \times N$ matrix $\mathbf{S}$ includes samples of the original source signals. $\mathbf{A}$ is an unknown regular $d \times d$ mixing matrix. It is assumed square because the number of mixtures and sources can always be made equal in this simple model.

In spite of the success of ICA in solving even large-scale real world problems, some theoretical questions remain partly open. One of the most central questions is the theoretical accuracy of the developed algorithms. Mostly the methods are compared through empirical studies, which may demonstrate the efficacy in various situations. However, the general validity cannot be proven like this. A natural question is, whether there exists some theoretical limit for separation performance, and whether it is possible to reach it.

Many of the algorithms can be shown to converge in theory to the correct solution giving the original sources, under the assumption that the sample size N is infinite. This is unrealistic. For finite data sets, what typically happens is that the sources are not completely unmixed but some traces of the other sources remain in them even after the algorithm has converged. This means that the obtained demixing matrix $\widehat{\mathbf{W}}$ is not exactly the inverse of $\mathbf{A}$, and the matrix of estimated sources $\mathbf{Y} = \widehat{\mathbf{W}}\mathbf{X} = \widehat{\mathbf{W}}\mathbf{A}\mathbf{S}$ is only approximately equal to $\mathbf{S}$. A natural measure of error is the deviation of the so-called gain matrix $\mathbf{G} = \widehat{\mathbf{W}}\mathbf{A}$ from the identity matrix, i.e., the variances of its elements.

The well-known lower limit for the variance of a parameter vector in estimation theory is the Cramér-Rao lower bound (CRB). In publications [1, 2], the CRB for the demixing matrix of the FastICA algorithm was derived. The result depends on the score functions of the sources,

$$\psi_k(s) = -\frac{d}{ds}\log p_k(s) = -\frac{p_k'(s)}{p_k(s)} \tag{2.2}$$

where $p_k(s)$ is the probability density function of the k-th source. Let

$$\kappa_k \quad = \quad \mathrm{E}\left[\psi_k^2(s_k)\right]. \tag{2.3}$$

Then, assuming that the correct score function is used as the nonlinearity in the FastICA algorithm, the asymptotic variances of the off-diagonal elements (k, ℓ) of matrix $\mathbf{G}$ for the one-unit and symmetrical FastICA algorithm, respectively, read

$$V_{k\ell}^{1U-opt} \quad = \quad \frac{1}{N}\frac{1}{\kappa_k - 1} \tag{2.4}$$

$$V_{k\ell}^{SYM-opt} \quad = \quad \frac{1}{N}\frac{\kappa_k + \kappa_\ell - 2 + (\kappa_\ell - 1)^2}{(\kappa_k + \kappa_\ell - 2)^2}, \tag{2.5}$$

while the CRB reads

$$\mathrm{CRB}(\mathbf{G}_{k\ell}) = \frac{1}{N}\frac{\kappa_k}{\kappa_k\kappa_\ell - 1}. \tag{2.6}$$

Comparison of these results implies that the algorithm FastICA is nearly statistically efficient in two situations:

(1) One-unit version FastICA with the optimum nonlinearity is asymptotically efficient for $\kappa_k \to \infty$, regardless of the value of κ_ℓ.

(2) Symmetric FastICA is nearly efficient for κ_i lying in a neighborhood of 1^+, provided that all independent components have the same probability distribution function, and the nonlinearity is equal to the joint score function.

The work was continued to find a version of the FastICA that would be asymptotically efficient, i.e. with some choice of nonlinearities would be able to attain the CRB. This work [3] will be reported later.

References

[1] Koldovský, Z., Tichavský, P. and Oja, E.: Cramér-Rao lower bound for linear independent component analysis. *Proc. Int. Conf. on Acoustics, Speech, and Signal Proc. (ICASSP'05)*, March 20 - 23, 2005, Philadelphia, USA (2005).

[2] Tichavský, P., Koldovský, Z. and Oja, E.: Performance analysis of the FastICA algorithm and Cramér-Rao bounds for linear independent component analysis. *IEEE Trans. on Signal Processing 54*, no. 4, April 2006.

[3] Koldovský, Z., Tichavský, P., and Oja, E.: Efficient variant of algorithm FastICA for independent component analysis attaining the Cramér-Rao lower bound. *IEEE Trans. on Neural Networks*, to appear (2006).

2.3 Nonlinear ICA and BSS

Juha Karhunen, Antti Honkela, Alexander Ilin

Recent advances in blind source separation (BSS) and independent component analysis (ICA) for nonlinear mixing models have been reviewed in the invited journal paper [1]. After a general introduction to BSS and ICA, uniqueness and separability issues are discussed in more detail, presenting some new results. A fundamental difficulty in the nonlinear BSS problem and even more so in the nonlinear ICA problem is that they provide non-unique solutions without extra constraints, which are often implemented by using a suitable regularization. In the paper [1], two possible approaches are explored in more detail. The first one is based on structural constraints. Especially, post-nonlinear mixtures are an important special case, where a nonlinearity is applied to linear mixtures. For such mixtures, the ambiguities are essentially the same as for the linear ICA or BSS problems. The second approach uses Bayesian inference methods for estimating the best statistical parameters, under almost unconstrained models in which priors can be easily added. In the later part of the paper [1], various separation techniques proposed for post-nonlinear mixtures and general nonlinear mixtures are reviewed.

Our own research on nonlinear BSS has concentrated on the Bayesian approach which is described in Sec. 4.4. The latest results include the use of kernel PCA to initialize the model for improved accuracy in highly nonlinear problems as well as a variational Bayesian generative model for post-nonlinear ICA.

There exist few comparisons of nonlinear ICA and BSS methods, and their limitations and preferable application domains have been studied only a little. We have experimentally compared two approaches introduced for nonlinear BSS: the Bayesian methods developed at the Neural Network Research Centre (NNRC) of Helsinki University of Technology, and the BSS methods introduced for the special case of post-nonlinear (PNL) mixtures developed at Institut National Polytechnique de Grenoble (INPG) in France. This comparison study took place within the framework of the European joint project BLISS on blind source separation and its applications.

The Bayesian method developed at NNRC for recovering independent sources consists of two phases: Applying the general nonlinear factor analysis (NFA) [3] to obtain Gaussian sources; and their further rotation with a linear ICA technique such as the FastICA algorithm [2]. The compared BSS method, developed at INPG for post-nonlinear mixtures, is based on minimization of the mutual information between the sources. It uses a separating structure consisting of nonlinear and linear stages [4].

Both approaches were applied to the same ICA problems with artificially generated post-nonlinear mixtures of two independent sources. Based on the experimental results, the following conclusions were drawn on the applicability of the INPG and Bayesian NFA+FastICA approaches to post-nonlinear blind source separation problems:

1. The INPG method performs better in classical post-nonlinear mixtures with the same number of sources and observations when all post-nonlinear distortions are invertible.

2. The performance of both methods can be improved by exploiting more mixtures than the number of sources especially in the case of noisy mixtures.

3. The advantage of the Bayesian methods in post-nonlinear BSS problems is that they can separate post-nonlinear mixtures with non-invertible post-nonlinearities

provided that the full mapping is globally invertible. The existing INPG methods cannot do this due to their constrained separation structure.

The results of this comparison study were presented in [5].

References

[1] C. Jutten and J. Karhunen. Advances in Blind Source Separation (BSS) and Independent Component Analysis (ICA) for nonlinear mixtures. *Int. J. of Neural Systems*, 14(5):267-292, 2004.

[2] A. Hyvärinen and E. Oja. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9(7):1483–1492, 1997.

[3] H. Lappalainen and A. Honkela. Bayesian nonlinear independent component analysis by multi-layer perceptrons. In Mark Girolami, editor, *Advances in Independent Component Analysis*, pages 93–121. Springer-Verlag, Berlin, 2000.

[4] A. Taleb and C. Jutten. Source separation in post-nonlinear mixtures. *IEEE Transactions on Signal Processing*, 47(10):2807–2820, 1999.

[5] A. Ilin, S. Achard, and C. Jutten. Bayesian versus constrained structure approaches for source separation in post-nonlinear mixtures. In *Proc. of the Int. Joint Conf. on Neural Networks (IJCNN 2004)*, pages 2181–2186, Budapest, Hungary, 2004.

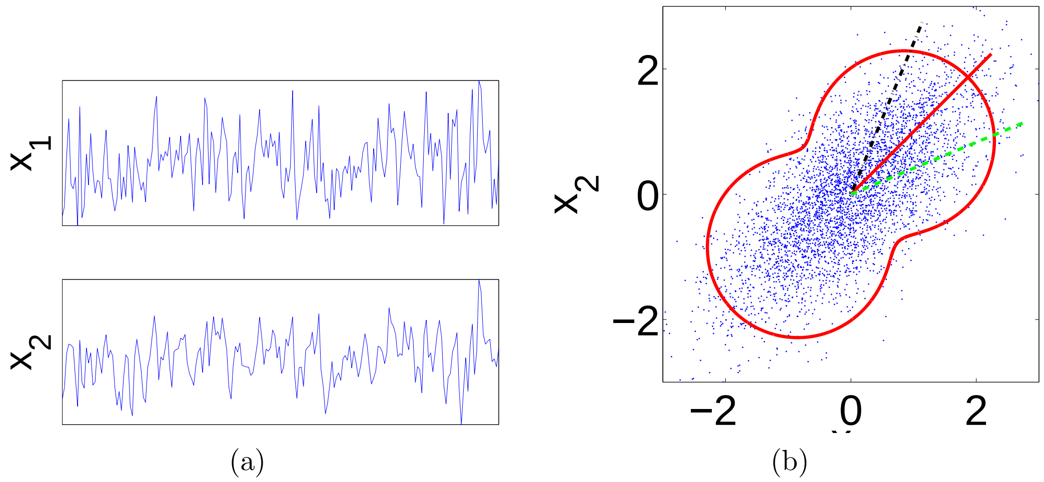
2.4 Denoising source separation

Jaakko Särelä

Denoising source separation (DSS,[1]) is a recently developed framework for linear source separation and feature extraction. In DSS, the algorithms are constructed around denoising operations. With certain types of denoising, DSS realises ICA.

With linear denoising, the algorithm consists of three steps: 1) sphering (whitening) 2) denoising 3) PCA. If the denoising is nonlinear, an iterative procedure has to be used, and the algorithm resembles nonlinear PCA and has the following iteration after presphering: 1) estimation of the sources using the current mapping, 2) denoising of the source estimates, 3) re-estimation of the mapping. The crucial part is the denoising, and available prior or acquired information may be implicitly implemented in it.

As an example, consider the two observations in Fig. 2.1. The correlation structure between the observations becomes apparent, when they are plotted against each others in a scatter-plot. The red curve illustrates the variance of different projections and the red line the direction where the variance is maximised. As it happens, the observations are linear mixtures of two independent sources. The mixing vectors are shown in the scatter-plot using the black and the green line.



(a) (b)

Figure 2.1: *a) Observed signals. b) The scatter-plot of the observed signals.*

As the original sources are independent of each others, a good attempt to recover them is to project the data to the principal components and thus remove any correlations between them. The resulting scatter-plot after normalisation of the variances (sphering or witening) is shown in Fig. 2.2. As illustrated by the red circle, the variance of any projection equals to one. However, the principal directions (y_1 and y_2) do not recover the original sources as shown by the black and the green line. Furthermore, there is no structure left in the scatter-plot.

The scatter-plot loses all the temporal structure the data may have. A good frequency representation is given by the discrete cosine transform (DFT). DFT of the sphered signals is shown in Fig. 2.3a. It seems that there are relatively more low frequencies than high frequencies. One hyphothesis could be that a source with low frequencies exists in the data. This source would become more visible (or denoised) by low-pass filtering. The resulting scatter-plot is shown in Fig. 2.3b.

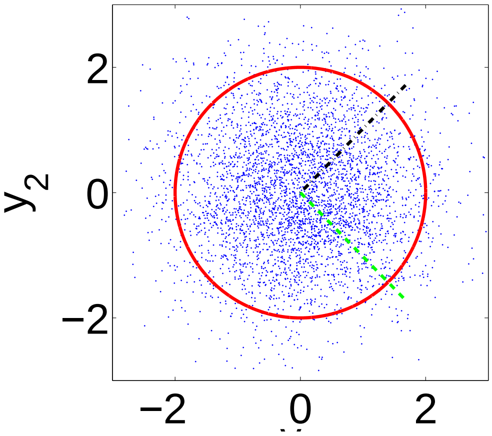
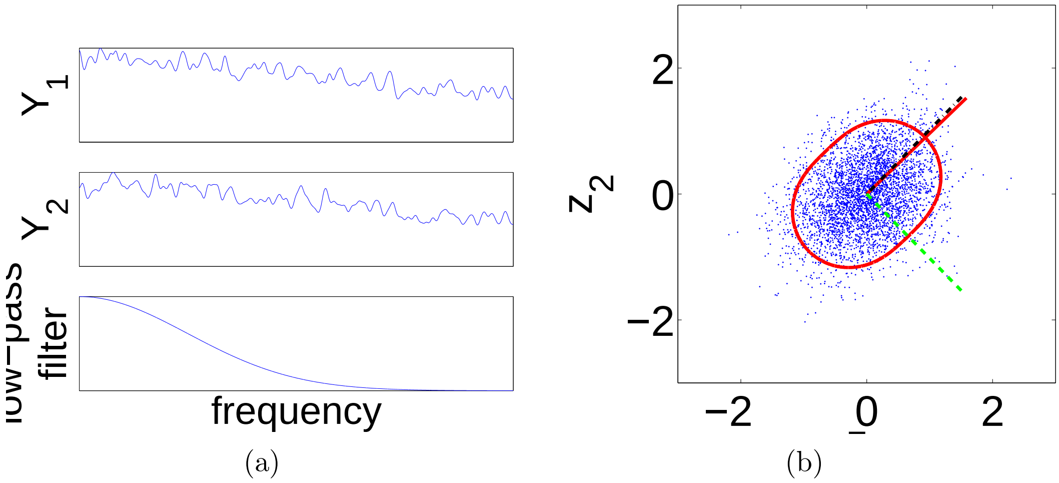


Figure 2.2: *Scatter-plot of the sphered signals.*



(a)

(b)

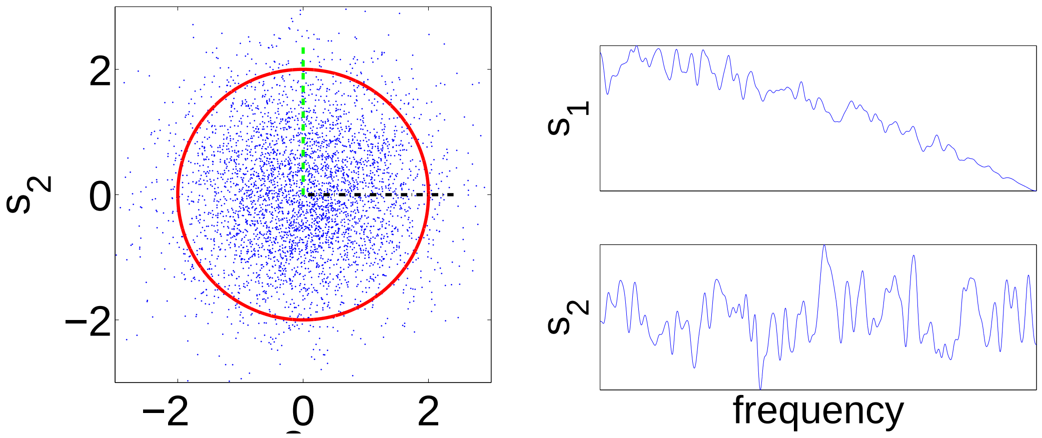
Figure 2.3: *a) Amplitude spectra of the sphered signals. b) The denoised signals.*

Now all directions do not have unit variance, and the maximal variance may be identified by another PCA. The principal directions align with the sphered mixing vectors (black and green line) and the original sources are recovered. The estimated sources and their amplitude spectra are shown in Fig. 2.4.

The DSS framework has been applied in several application fields. In this laboratory, we have applied it, for instance, to blind suppression of various interfering signals appearing in direct sequence CDMA communication systems (Sec. 2.6), to exploratory source separation of climate phenomena (Sec. 2.5) and to neuroinformatics (Ch. 3).

References

[1] J. Särelä and H. Valpola, “Denoising source separation,” *Journal of Machine Learning Research*, vol. 6, pp. 233–272, 2005.



(a) (b)

Figure 2.4: *a) The estimated sources. b) The amplitude spectra of the estimated sources.*

2.5 Climate data analysis with DSS

Alexander Ilin, Harri Valpola, Erkki Oja

One of the main goals of statistical analysis of climate data is to extract physically meaningful patterns of climate variability from highly multivariate weather measurements. The classical technique for defining such dominant patterns is principal component analysis (PCA), or empirical orthogonal functions (EOF) as it is called in climatology (see, e.g., [1]). However, the maximum remaining variance criterion used in PCA can lead to such problems as mixing different physical phenomena in one extracted component [2]. This makes PCA a useful tool for information compression but limits its ability to isolate individual modes of climate variation.

To overcome this problem, rotation of the principal components has proven useful. The classical rotation criteria used in climatology are based on the general concept of "simple structure" which can provide spatially or temporally localizated components [2]. Denoising source separation (DSS) is a tool which can also be used for rotating the principal components. It is particularly efficient when some prior information exists (e.g., the general shape of the time curves of the sources or their frequency contents). For example, in the climate data analysis we might be interested in some phenomena that would be cyclic over a certain period, or exhibit slow changes. Then, exploiting the prior knowledge may significantly help in finding a good representation of the data.

We use the DSS framework for exploratory analysis of the large spatio-temporal dataset provided by the NCEP/NCAR reanalysis project [3]. The data is the reconstruction of the daily weather measurements around the globe for a period of 56 years.

In our first works, we concentrate on slow climate oscillations and analyze three major atmospheric variables: surface temperature, sea level pressure and precipitation. In [4], we show that optimization of the criterion that we term clarity helps find the sources exhibiting the most prominent periodicity in a specific timescale. In the experiments, the components with the most prominent interannual oscillations are clearly related to the well-known El Niño–Southern Oscillation (ENSO) phenomenon. For all three variables, the most prominent component is a good ENSO index (see Fig. 2.5–2.6) and the second component is close to the derivative of the first one.

In [5], we extend the analysis to a more general case where slow components are separated by their frequency contents. The sources found using the frequency-based criterion give a meaningful representation of the slow climate variability as combination of trends, interannual oscillations, the annual cycle and slowly changing seasonal variations.

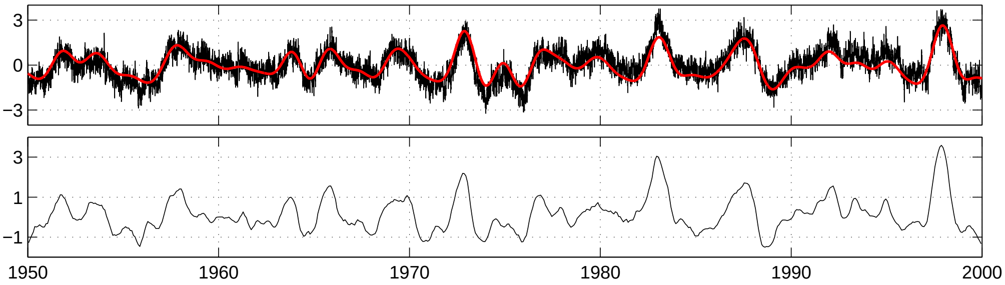


Figure 2.5: The dark curve on the upper plot shows the component with the most prominent interannual oscillations extracted with DSS. The red curve is the found component filtered in the interannual timescale. The lower plot presents the index which is used in climatology to measure the strength of El Niño. The curves have striking resemblance.

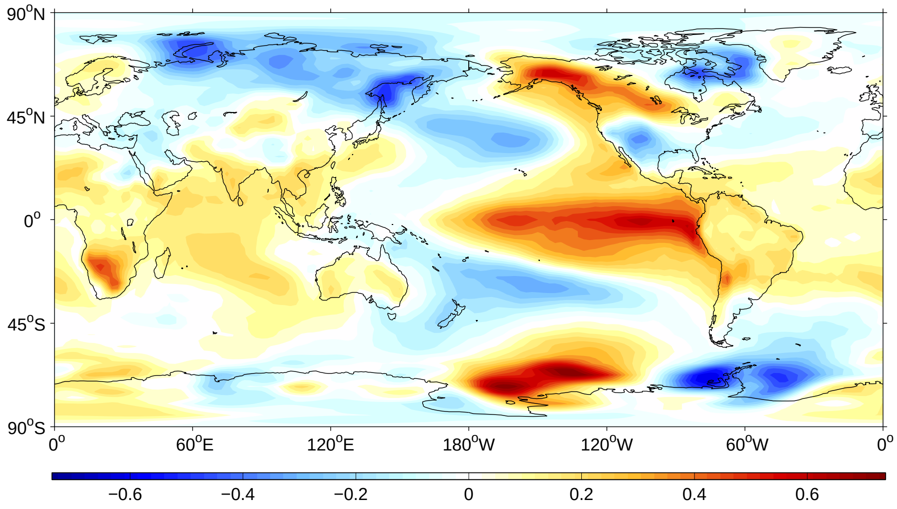


Figure 2.6: The temperature pattern corresponding to the component with the most prominent interannual oscillations. The map tells how strongly the component is expressed in the measurement data. The pattern has many features traditionally associated with El Niño. The scale of the map is in degrees centigrade.

References

[1] H. von Storch, and W. Zwiers. *Statistical Analysis in Climate Research.* Cambridge University Press, Cambridge, U.K, 1999.

[2] M. B. Richman. Rotation of principal components. *Journal of Climatology*, 6:293–335, 1986.

[3] E. Kalnay and coauthors. The NCEP/NCAR 40-year reanalysis project. Bulletin of the American Meteorological Society, 77:437–471, 1996.

[4] A. Ilin, H. Valpola, and E. Oja. Semiblind source separation of climate data detects El Niño as the component with the highest interannual variability. In *Proc. of the Int. Joint Conf. on Neural Networks (IJCNN 2005)*, pages 1722–1727, Montréal, Québec, Canada, 2005.

[5] A. Ilin, and H. Valpola. Frequency-based separation of climate signals. In *Proc. of 9th European Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD 2005)*, pages 519–526, Porto, Portugal, 2005.

2.6 ICA and denoising source separation in CDMA communications

**Karthikesh Raju, Tapani Ristaniemi, Juha Karhunen,
Jaakko Särelä and Erkki Oja**

In wireless communication systems, like mobile phones, an essential issue is division of the common transmission medium among several users. A primary goal is to enable each user of the system to communicate reliably despite the fact that the other users occupy the same resources, possibly simultaneously. As the number of users in the system grows, it becomes necessary to use the common resources as efficiently as possible.

During the last years, various systems based on CDMA (Code Division Multiple Access) techniques [1, 2] have become popular, because they offer several advantages over the more traditional FDMA and TDMA schemes based on the use of non-overlapping frequency or time slots assigned to each user. Their capacity is larger, and it degrades gradually with increasing number of simultaneous users who can be asynchronous. On the other hand, CDMA systems require more advanced signal processing methods, and correct reception of CDMA signals is more difficult because of several disturbing phenomena [1, 2] such as multipath propagation, possibly fading channels, various types of interferences, time delays, and different powers of users.

Direct sequence CDMA data model can be cast in the form of a linear independent component analysis (ICA) or blind source separation (BSS) data model [3]. However, the situation is not completely blind, because there is some prior information available. In particular, the transmitted symbols have a finite number of possible values, and the spreading code of the desired user is known.

In this project, we have applied independent component analysis and denoising source separation (DSS) to blind suppression of various interfering signals appearing in direct sequence CDMA communication systems. The standard choice in communications for suppressing such interfering signals is the well-known RAKE detection method [2]. RAKE utilizes available prior information, but it does not take into account the statistical independence of the interfering and desired signal. On the other hand, ICA utilizes this independence, but it does not make use of the prior information. Hence it is advisable to combine the ICA and RAKE methods for improving the quality of interference cancellation.

In the journal paper [4], various schemes combining ICA and RAKE are introduced and studied for different types of interfering jammer signals under different scenarios. By using ICA as a preprocessing tool before applying the conventional RAKE detector, some improvement in the performance is achieved, depending on the signal-to-interference ratio, signal-to-noise ratio, and other conditions [4]. These studies have been extended to consider multipath propagation and coherent jammers in [5].

All these ICA-RAKE detection methods use the FastICA algorithm [3] for separating the interfering jammer signal and the desired signal. In the case of multipath propagation, it is meaningful to examine other temporal separation methods, too. The results of such a study have been presented in [7].

The paper [6] deals with application of denoising source separation [9] to interference cancellation. This is a semi-blind approach which uses the spreading code of the desired user but does not require training sequences. The results of the DSS-based interference cancellation scheme show improvements over conventional detection.

Work on both uplink and downlink interference cancellation in direct sequence CDMA systems has been summarized in the joint paper [8]. In this paper, an effort is made to present both uplink and downlink methods under a unified framework.

References

[1] S. Verdu, *Multiuser Detection*. Cambridge Univ. Press, 1998.

[2] J. Proakis, *Digital Communications*. McGraw-Hill, 3rd edition, 1995.

[3] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. Wiley, 2001, 481+xxii pages.

[4] K. Raju, T. Ristaniemi, J. Karhunen, and E. Oja, Jammer cancellation in DS-CDMA arrays using independent component analysis. *IEEE Trans. on Wireless Communications*, Vol. 5, No. 1, January 2006, pp. 77–82.

[5] K. Raju, T. Ristaniemi, and J. Karhunen, Semi-blind interference suppression on coherent multipath environments. In *Proc. of the First IEEE Int. Symp. of Control, Communications, and Signal Processing (ISCCSP2004)*, Hammamet, Tunisia, March 21-24, 2004, pp. 283–286.

[6] K. Raju and J. Särelä, A denoising source separation based approach to interference cancellation for DS-CDMA array systems. In *Proc. of the 38th Asilomar Conf. on Signals, Systems, and Computers*, Pacific Grove, California, USA, November 7-10, 2004, pp. 1111-1114.

[7] K. Raju and B. Phani Sudheer, Blind source separation for interference cancellation - a comparison of several spatial and temporal statistics based techniques. In *Proc. of the 3rd Workshop on the Internet, Telecommunications, and Signal Processing*, Adelaide, Australia, December 20-22, 2004.

[8] K. Raju, T. Huovinen, and T. Ristaniemi, Blind interference cancellation schemes for DS-CDMA systems. In *Proc. of the IEEE Int. Symp. on Antennas and Propagation and UNSC/URSI National Radio Science Meeting*, Washington, USA, July 3-8, 2005.

[9] J. Särelä and H. Valpola, Denoising source separation. *J. of Machine Learning Research*, Vol. 6, 2005, pp. 233–272.

2.7 ICA for image representations

Mika Inki

Already the earliest adapters of ICA on small image windows noted the similarity of the features to cortical simple cell receptive fields [1, 5]. This can be considered as support for the notion that the primary visual cortex (and early visual system in general) employs a strategy of sparse coding or redundancy reduction. In any case, the features obtained by ICA, and especially their efficiency in image coding and functionality in edge detection, can be argued to be useful when the objective is to build a hierarchical system capable of image analysis or understanding.

However, there are many limitations on the usefulness of the ICA description of images. A basic limitation is that ICA considers the components to be independent, which they are not in any sense with image data. Also, it can be argued that every possible scaling, translation and rotation of every ICA feature should also be in the basis, resulting in very highly overcomplete description, computationally infeasible to estimate. Another computational hindrance is the small window size necessitated by the curse of dimensionality.

We have focused on removing these limitations, and extending the ICA model to better account for image statistics, while comparing it to biological visual systems. We have, for example, examined the dependencies between ICA features in image data [3], built models based on these findings, studied overcomplete models [2, 4], and examined how the features can be extended past the window edges, cf. Figure 2.7.

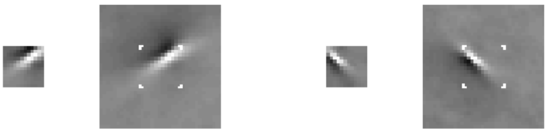


Figure 2.7: A couple of typical ICA features for images and their extensions.

References

[1] A.J. Bell and T.J. Sejnowski. The ’independent components’ of natural scenes are edge filters. *Vision Research*, 37:3327–3338, 1997.

[2] A. Hyvärinen and M. Inki. Estimating overcomplete independent component bases for image windows. *Journal of Mathematical Imaging and Vision*, 2002.

[3] M. Inki. A model for analyzing dependencies between two ICA features in natural images. In *Proc. Int. Conf. on Independent Component Analysis and Blind Signal Separation*, Granada, Spain, 2004.

[4] M. Inki. An easily computable eight times overcomplete ICA method for image data. In *Proc. Int. Conf. on Independent Component Analysis and Blind Signal Separation*, Charleston, South Carolina, 2006.

[5] B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37:3311–3325, 1997.

2.8 Analyzing 0-1 data

Ella Bingham

A novel probabilistic latent variable method for analyzing 0-1-valued (binary) data was presented. This method, termed as "aspect Bernoulli", was first described in [1]. The method is able to detect and distinguish between two types of 0s in the data: those that are "true absences" and those that are "false absences"; both of these are coded as 0 in the data.

As an example we may consider text documents in which some words are missing because they do not fit the topical content of the data — they are "true absences". Some other words are missing just because the author of the document did not use them although they would nicely fit the topic — these are "false absences" and the document could be augmented with these words. Another application might be black-and-white images in which some pixels are turned to white by an external noise process, resulting in "true" and "false" white pixels. Our method finds a specific latent component that accounts for the "false absences". Figure 2.8 shows results on this.

Similarly, the method can distinguish between two types of 1s: "true presences" and "false presences"; the latter could be extra black pixels in a black-and-white image, for example.

The method can be used in several applications: noise removal in black-and-white images; detection of false data instances in noisy data; and query expansion where topically related words are added into a document.

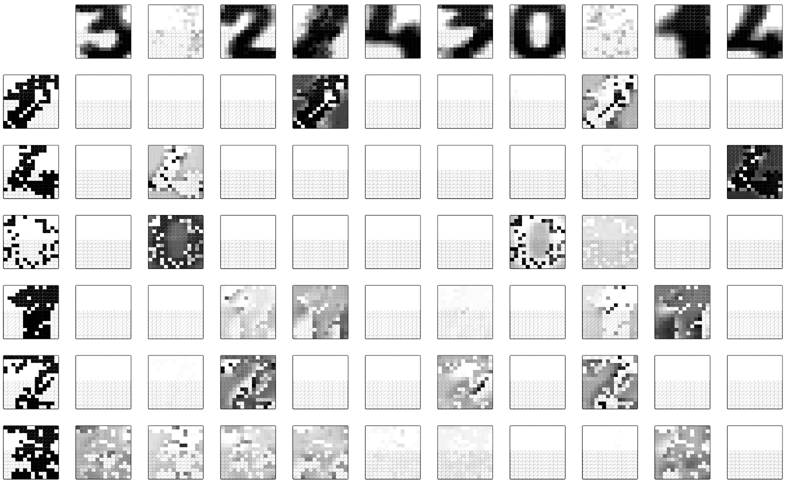


Figure 2.8: Analyzing corrupted black-and-white images. The top row shows the basis images estimated by the aspect Bernoulli model when corrupted images are fed into the algorithm. Examples of observed corrupted images are shown in the first column. The middle rows and columns give the pixel-wise probability that a basis image is responsible for generating a pixel in the observed image. For example, the observed digit "1" in the second row is largely generated by the 4th basis image resembling the digit "1", but the corrupted pixels are generated by the 8th basis image which is almost completely white and accounts for the corruption.

References

[1] Ata Kabán, Ella Bingham and Teemu Hirsimäki. Learning to read between the lines: The aspect Bernoulli model. *Proceedings of the 4th SIAM International Conference on Data Mining*, pp.462–466, 2004.

Chapter 3

Neuroinformatics

Ricardo Vigário, Jaakko Särelä, Sergey Borisov, Jarkko Ylipaavalniemi, Antti Honkela, Alexander Ilin, Elina Karp, Erkki Oja

3.1 Setup of the group

In the period spanned by this report, a neuroinformatics group was formally established. This field of research has been defined as *the combination of neuroscience and information sciences to develop and apply advanced tools and approaches essential for a major advancement in understanding the structure and function of the brain.* Aside from the development of the tools, this often means that the fields of application include the analysis and modelling of neuronal behaviour, as well as the efficient handling and mining of scientific databases.

From a methodological viewpoint, the neuroinformatics group has been involved in studying certain properties of ICA, such as its reliability and applicability to the analysis of electrophysiological brain data (namely electroencephalograms, EEGs and magnetoencephalograms, MEG). Also the denoising source separation framework (DSS) was introduced, in the more general context of linear source separation and feature extraction (see Sec.2.4). From a biomedical signal processing viewpoint, we have shown the usability of DSS to study MEG signals, as is illustrated in Sec.3.3. Also using DSS, we investigated different possible origins for high- and low-amplitude alpha-activity in EEG (see Sec.3.2).

Several other topics have been researched in the field of biomedical signal processing, which are not thoroughly reported here. These lines of research will only have a visible outcome in the next biennial report. They include the analysis of synchronous activity in the brain; a more thorough study on the applicability of ICA to MEG; as well as the application of our methods of tissue segmentation in magnetic resonance imaging (MRI), to the detection of brain lesions. We also continued our research on technical aspects of the analysis of functional magnetic resonance imaging (fMRI), and progress towards the analysis of more natural stimuli.

References

[1] J. Särelä and H. Valpola, “Denoising source separation,” *Journal of Machine Learning Research*, vol. 6, pp. 233–272, 2005.

[2] Borisov, S., A. Ilin, R. Vigário, A. Kaplan, and E. Oja. ”Does low- and high-amplitude segments of alpha activity belong to different sources: studying of EEG using both segmental analysis and DSS?” In *Proc. of 11th Annual Meeting of the Organization for Human Brain Mapping*, Toronto, Canada, 2005.

[3] Honkela, A., T. Östman, and R. Vigário. ”Empirical evidence of the linear nature of magnetoencephalograms”. In *Proc. of 13th European Symposium on Artificial Neural Networks (ESANN’05)*, Bruges, Belgium, 2005.

[4] Müller, K.-R., R. Vigário, F. Meinecke, and A. Ziehe. ”Blind source separation techniques for decomposing event related brain signals”. *Int. Journal of Bifurcation and Chaos 14*(2), 773–792, 2004.

[5] Karp, E., H. Gävert, J. Särelä, and R. Vigário. ”Independent component analysis decomposition of structural MRI”. In *Proc. of 2nd IASTED Int. Conf. on Biomed. Eng. (BioMED’04)*, Innsbruck, Austria, 2004.

[6] Karp, E. and R. Vigário. ”Unsupervised MRI tissue classification by support vector machines”. In *Proc. of 2nd IASTED Int. Conf. on Biomed. Eng. (BioMED’04)*, Innsbruck, Austria, 2004.

[7] Ylipaavalniemi, J., and R. Vigário. "Analysis of Auditory fMRI Recordings via ICA: A Study on Consistency". In *Proc. of Int. Joint Conf. on Neural Networks (IJCNN'04)*, Budapest, Hungary, 2004.

[8] Raitio, J., R. Vigário, J. Särelä, and T. Honkela. "Assessing similarity of emergent representations based on unsupervised learning". In *Proc. of Int. Joint Conf. on Neural Networks (IJCNN'04)*, Budapest, Hungary, 2004.

3.2 Source localization of low- and high-amplitude alpha activity

In this work, we addressed the question of whether low- and high-amplitude alpha activities have common brain activation sources. Due to the well-documented nonstationary nature of electroencephalograms (EEG), we could not apply our methods directly to the complete data set. Yet, in practice, it can be considered as a sequence of quasi-stationary segments. Therefore, we used a segmental analysis in all our processing. This was used to segregate between the higher and lower alpha amplitude segments. In addition, we used the denoising source separation (DSS) method to isolate EEG components in the alpha range.

We analyzed EEGs of 9 healthy subjects. We used 16 electrodes, in the standard 10/20 configuration. Subjects were asked to rest with their eyes closed. The segmentation analysis of alpha activity was performed for each subject separately. Then, segmented EEGs were divided in two sequences, comprising the 50% high- and 50% low-amplitude segments, respectively (see Fig. 3.1).

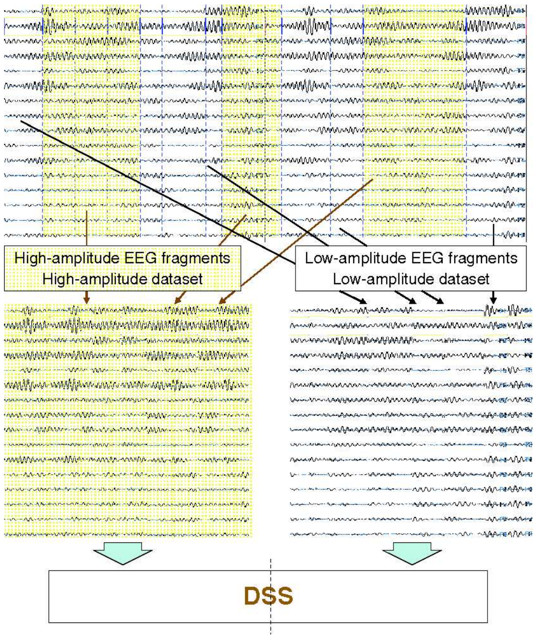


Figure 3.1: Original EEG data for one subject. Pre-processing segmentation of the high- and low-amplitude alpha activity, based on a reference channel over the occipital area.

The DSS algorithm was then applied to both sequences separately, yielding two sets of sources. The spatial patterns associated with these sources, for both high- and low-amplitude segments, were analyzed using standard k-means cluster analysis. We observe that most clustering procedures displayed similar spatial patterns of sources across different subjects, suggesting the existence of reliable sets of neural sources of alpha activity across subjects. Also we see (Fig. 3.2), that some clusters discriminate between high- and low-amplitude segments. This suggests that each neuronal source is mostly involved in either high or low amplitude activity (respectively depicted as yellow and clear frames).

According to the results, it is possible to assume that both low- and high-amplitude alpha activity may have both common and distinctive bioelectrical sources in the brain.

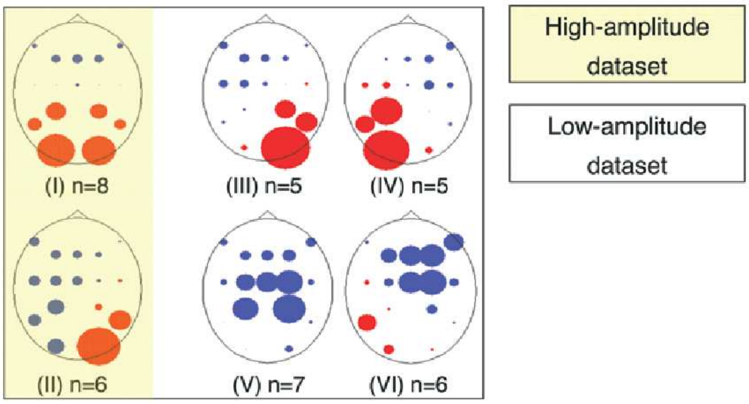


Figure 3.2: Averaged spatial patterns of the sources within the clusters. n is the number of sources in the clusters. All sources within the same cluster belong to different subjects (of 9). Frame I and II correspond to high-amplitude alpha, whereas the others display low.

The different spatial patterns of sources found for low- and high-amplitude may suggest that these populations of alpha-activity have their own nature and could perform different physiological functions.

3.3 DSS extraction of the cardiac subspace in MEG

In this section, we demonstrate the capability of the denoising source separation framework (DSS, see Sec.2.4) to extract some very weak cardiac signals, using detailed prior information in an adaptive manner.

Denoising of the cardiac subspace

A clear QRS complex, which is the main electromagnetic pulse in the cardiac cycle, can be extracted from the MEG data using standard BSS methods, such as kurtosis- or tanh-based denoising. Due to its sparse nature, this QRS signal can be used to estimate the places of the heart beats. With the places known, we can guide further search using the averaging DSS, discussed below.

Consider the source estimate in Fig. 3.3a. Let us assume to be known beforehand that the signal has a repetitive structure and that the average repetition rate is known. The quasi-periodicity of the signal can be used to perform DSS to get a better estimate. The denoising proceeds as follows:

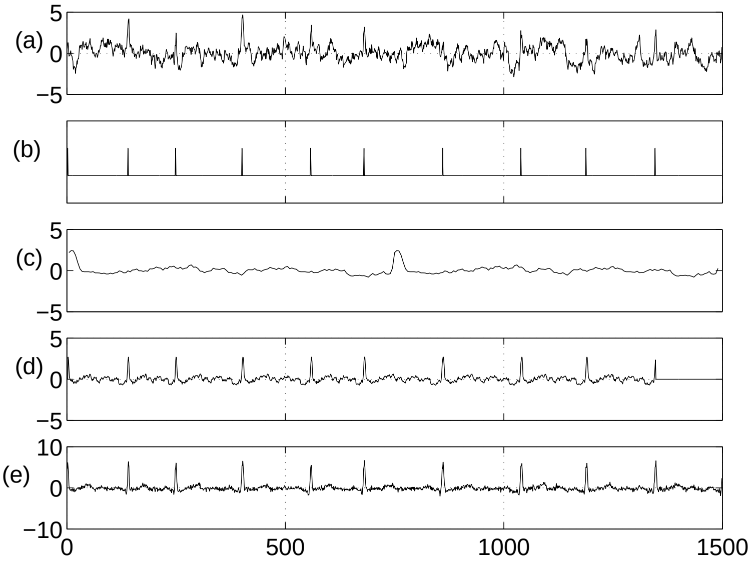


Figure 3.3: *a) Current source estimate of a quasiperiodic signal b) Peak estimates c) Average signal (two periods are shown for clarity). d) Denoised source estimate. e) Source estimate corresponding to the re-estimated.*

1. Estimate the locations of the peaks of the current source estimate (Fig. 3.3b).
2. Chop each period from peak to peak.
3. Dilate each period to a fixed length L (linearly or nonlinearly).
4. Average the dilated periods (Fig. 3.3c).
5. Let the denoised source estimate be a signal where each period has been replaced by the averaged period dilated back to its original length (Fig. 3.3d).

The re-estimated signal in Fig. 3.3e, based on the denoised signal, shows significantly better SNR compared to the original source estimate, in Fig. 3.3a.

When the estimation of the QRS locations has been stabilised, a subspace that is composed of signals having activity phase-locked to the QRS complexes can be extracted.

Separation results

Figure 3.4 depicts five signals averaged around the QRS complexes, found using the procedure above[1]. The first signal presents a very clear QRS complex, whereas the second one contains the small P and the T waves. An interesting phenomenon is found in the third signal: there is a clear peak at the QRS onset, which is followed by a slow attenuation phase. We presume that it originates from some kind of relaxing state.

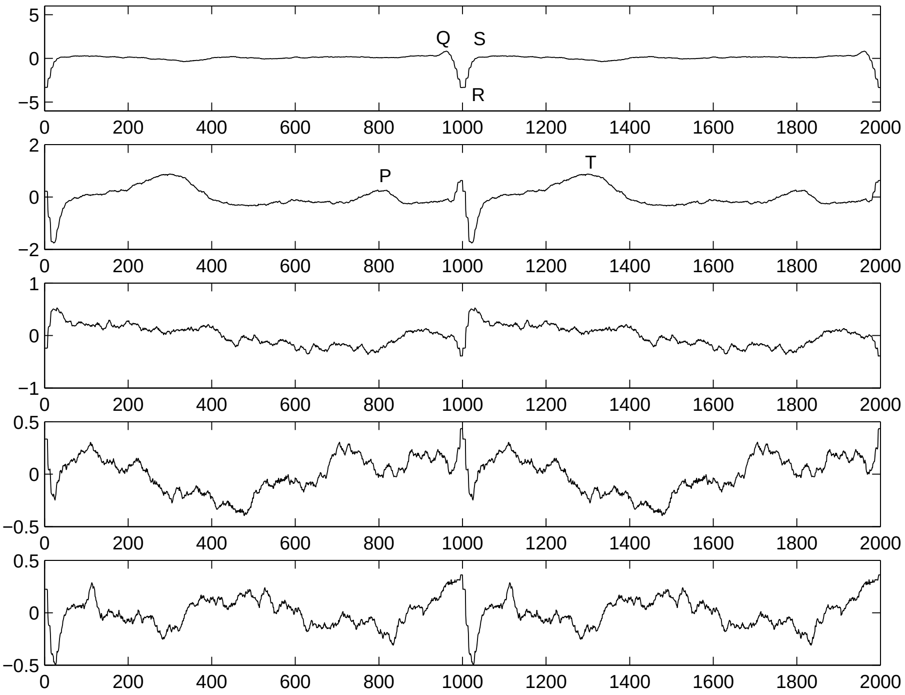


Figure 3.4: *Averages of three heart-related signals and presumably two overfitting results.*

Two other heart-related signals were also extracted. They both show a clear deflection during the QRS complex, but have as well significant activity elsewhere. These two signals might present a case of overfitting. It is worth noticing that even the strongest component of the cardiac subspace is rather weakly present in the original data. The other components of the subspace are hardly detectable without advanced methods beyond blind source separation. This clearly demonstrates the power that DSS can provide for an exploring researcher.

[1] For clarity, two identical cycles of averaged heart beats are always shown.

Chapter 4

Variational Bayesian learning of generative models

Juha Karhunen, Antti Honkela, Alexander Ilin, Tapani Raiko, Markus Harva, Harri Valpola, Erkki Oja

4.1 Bayesian modeling and variational learning: introduction

Unsupervised learning methods are often based on a generative approach where the goal is to find a model which explains how the observations were generated. It is assumed that there exist certain source signals (also called factors, latent or hidden variables, or hidden causes) which have generated the observed data through an unknown mapping. The goal of generative learning is to identify both the source signals and the unknown generative mapping.

The success of a specific model depends on how well it captures the structure of the phenomena underlying the observations. Various linear models have been popular, because their mathematical treatment is fairly easy. However, in many realistic cases the observations have been generated by a nonlinear process. Unsupervised learning of a nonlinear model is a challenging task, because it is typically computationally much more demanding than for linear models, and flexible models require strong regularization.

In Bayesian data analysis and estimation methods, all the uncertain quantities are modeled in terms of their joint probability distribution. The key principle is to construct the joint posterior distribution for all the unknown quantities in a model, given the data sample. This posterior distribution contains all the relevant information on the parameters to be estimated in parametric models, or the predictions in non-parametric prediction or classification tasks [1].

Denote by $\mathcal{H}$ the particular model under consideration, and by $\boldsymbol{\theta}$ the set of model parameters that we wish to infer from a given data set X. The posterior probability density $p(\boldsymbol{\theta}|X,\mathcal{H})$ of the parameters given the data X and the model $\mathcal{H}$ can be computed from the Bayes' rule

$$p(\boldsymbol{\theta}|X,\mathcal{H}) = \frac{p(X|\boldsymbol{\theta},\mathcal{H})p(\boldsymbol{\theta}|\mathcal{H})}{p(X|\mathcal{H})} \tag{4.1}$$

Here $p(X|\boldsymbol{\theta},\mathcal{H})$ is the likelihood of the parameters $\boldsymbol{\theta}$, $p(\boldsymbol{\theta}|\mathcal{H})$ is the prior pdf of the parameters, and $p(X|\mathcal{H})$ is a normalizing constant. The term $\mathcal{H}$ denotes all the assumptions made in defining the model, such as choice of a multilayer perceptron (MLP) network, specific noise model, etc.

The parameters $\boldsymbol{\theta}$ of a particular model $\mathcal{H}_i$ are often estimated by seeking the peak value of a probability distribution. The non-Bayesian maximum likelihood (ML) method uses to this end the distribution $p(X|\boldsymbol{\theta},\mathcal{H})$ of the data, and the Bayesian maximum a posteriori (MAP) method finds the parameter values that maximize the posterior probability density $p(\boldsymbol{\theta}|X,\mathcal{H})$. However, using point estimates provided by the ML or MAP methods is often problematic, because the model order estimation and overfitting (choosing too complicated a model for the given data) are severe problems [1].

Instead of searching for some point estimates, the correct Bayesian procedure is to use all possible models to evaluate predictions and weight them by the respective posterior probabilities of the models. This means that the predictions will be sensitive to regions where the probability mass is large instead of being sensitive to high values of the probability density [2]. This procedure optimally solves the issues related to the model complexity and choice of a specific model $\mathcal{H}_i$ among several candidates. In practice, however, the differences between the probabilities of candidate model structures are often very large, and hence it is sufficient to select the most probable model and use the estimates or predictions given by it.

A problem with fully Bayesian estimation is that the posterior distribution (4.1) has a highly complicated form except for in the simplest problems. Therefore it is too difficult to handle exactly, and some approximative method must be used. Variational methods form a class of approximations where the exact posterior is approximated with a simpler distribution [3]. In a method commonly known as *Variational Bayes (VB)* [1, 2] the misfit of the approximation is measured by the Kullback-Leibler (KL) divergence between two probability distributions $q(v)$ and $p(v)$. The KL divergence is defined by

$$D(q \parallel p) = \int q(v) \ln \frac{q(v)}{p(v)} dv \tag{4.2}$$

which measures the difference in the probability mass between the densities $q(v)$ and $p(v)$.

A key idea in the VB method is to minimize the misfit between the actual posterior pdf and its parametric approximation using the KL divergence. The approximating density is often taken a diagonal multivariate Gaussian density, because the computations become then tractable. Even this crude approximation is adequate for finding the region where the mass of the actual posterior density is concentrated. The mean values of the Gaussian approximation provide reasonably good point estimates of the unknown parameters, and the respective variances measure the reliability of these estimates.

A main motivation of using VB is that it avoids overfitting which would be a difficult problem if ML or MAP estimates were used. VB method allows one to select a model having appropriate complexity, making often possible to infer the correct number of sources or latent variables. It has provided good estimation results in the very difficult unsupervised (blind) learning problems that we have considered.

Variational Bayes is closely related to information theoretic approaches which minimize the description length of the data, because the description length is defined to be the negative logarithm of the probability. Minimal description length thus means maximal probability. In the probabilistic framework, we try to find the sources or factors and the nonlinear mapping which most probably correspond to the observed data. In the information theoretic framework, this corresponds to finding the sources and the mapping that can generate the observed data and have the minimum total complexity. The information theoretic view also provides insights to many aspects of learning and helps explain several common problems [4].

In the following subsections, we first present some recent theoretical improvements to VB methods and a practical building block framework that can be used to easily construct new models. After this we discuss practical models for nonlinear and non-negative blind source separation as well as multivariate time series analysis using nonlinear state-space models. A more structured extension of probabilistic relational models is also presented. Finally we present applications of the developed Bayesian methods to astronomical data analysis problems.

4.2 Theoretical improvements

Effect of posterior approximation

Most applications of variational Bayesian learning to ICA models reported in the literature assume a fully factorized posterior approximation $q(v)$, because this usually results in a computationally efficient learning algorithm. However, the simplicity of the posterior approximation does not allow for representing all possible solutions, which may greatly affect the found solution.

Our paper [5] shows that neglecting the posterior correlations of the sources $\mathbf{S}$ in the approximating density $q(\mathbf{S})$ introduces a bias in favor of the principal component analysis (PCA) solution. By the PCA solution we mean the solution which has an orthogonal mixing matrix. Nevertheless, if the true mixing matrix is close to orthogonal and the source model is strongly in favor of the desirable ICA solution, the optimal solution can be expected to be close to the ICA solution. In [5], we studied this problem both theoretically and experimentally by considering linear ICA models with either independent dynamics or non-Gaussian source models. The analysis also extends to the case of nonlinear mixtures.

Figure 4.1 presents experimental results illustrating the general trade-off of variational Bayesian learning between the misfit of the posterior approximation and the accuracy of the model. According to our assumption, the sources can be accurately modeled in the ICA solution and therefore the cost of inaccurate assumption would increase towards the ICA solution. As a result, the ICA solution is found for strongly non-Gaussian sources ($\nu = 1$). On the other hand, if the true mixing matrix is not orthogonal, the optimal posterior covariance of the sources could have posterior correlations between the sources. Then, the misfit of the posterior approximation of the sources is minimized in the PCA solution where the true posterior covariance would be diagonal. This is the reason why the PCA solution is found for the sources whose distribution is close to Gaussian ($\nu = 0.6$). In the intermediate cases ($\nu = 0.7, \nu = 0.9$), some compromise solutions, which lie in between the PCA and ICA solutions, can be found.

Accurate linearisation for learning nonlinear models

Learning of nonlinear models in the variational Bayesian framework fundamentally reduces to evaluating statistics of the data predicted by the model as a function of the parameters of the variational approximation of the posterior distribution. This is equivalent to evaluating statistics of a nonlinear transformation of the approximating probability distribution. A common approach that was also used in our earlier work on nonlinear models [6, 7] is to use a Taylor series approximation to linearise the nonlinearity. Unfortunately this approximation breaks down when the variance of the approximating distribution increases, and this leads to algorithmic instability.

For handling this problem, a new linearisation method based on replacing the local approach of the Taylor scheme with a more global approximation was proposed in [8, 9]. In case of multilayer perceptron (MLP) networks this can be done efficiently by replacing the nonlinear activation function of the hidden neurons by a linear function that would provide the same output mean and variance, as evaluated by Gauss–Hermite quadrature. The resulting approximation yields significantly more accurate estimates of the cost of the model while being computationally almost as efficient. This is illustrated in Figure 4.2.

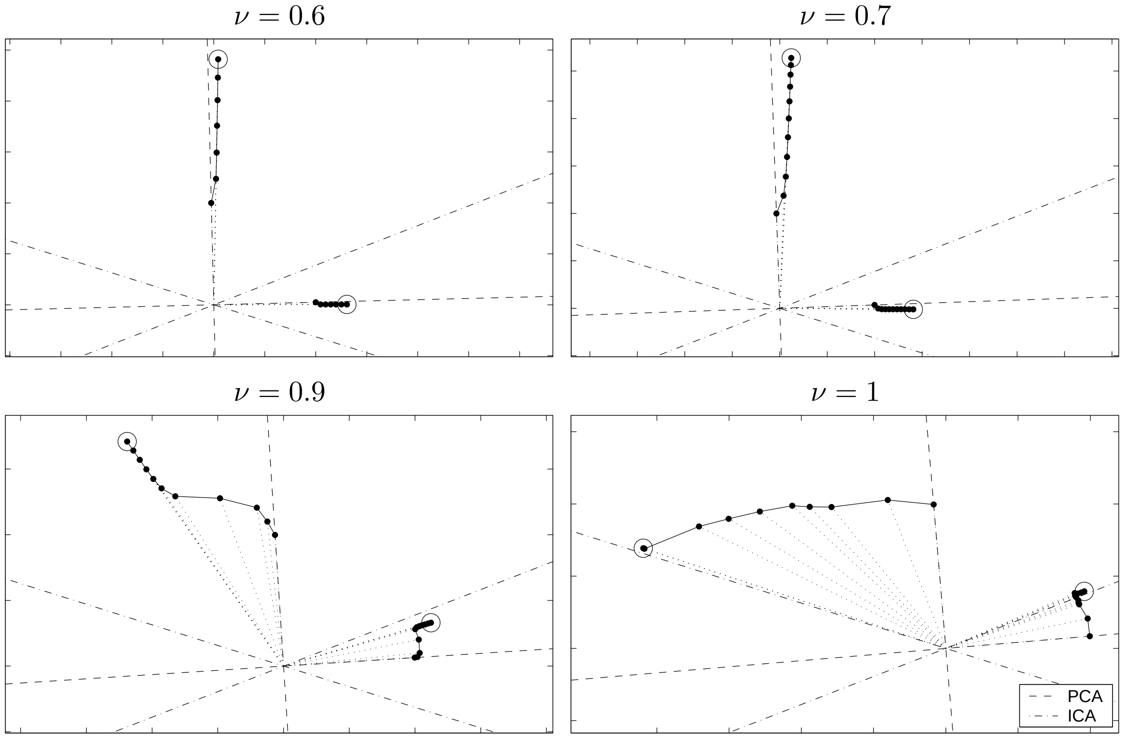


Figure 4.1: Separation results obtained with a model with super-Gaussian sources and fully factorial approximation for four test ICA problems. The parameter ν is the measure of the non-Gaussianity of the sources used in the test data. The dotted lines represent the columns of the mixing matrix during learning, the final solution is circled. The PCA and ICA directions are shown on the plots with the dashed and dashed-dotted lines respectively.

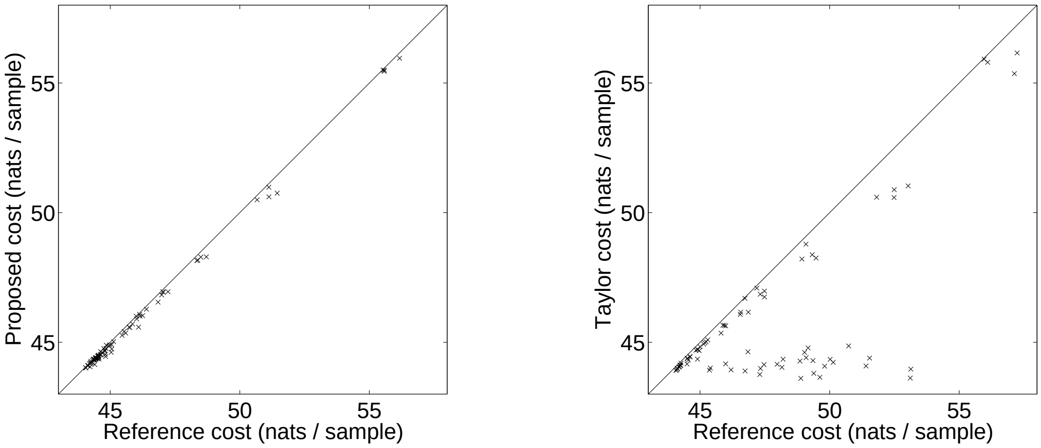


Figure 4.2: The attained values of the cost in different simulations as evaluated by the different approximations plotted against reference values evaluated by sampling. The left subfigure shows the values from experiments using the proposed approximation and the right subfigure from experiments using the Taylor approximation.

Partially observed values

It is well known that Bayesian methods provide well-founded and straightforward means for handling missing values in data. The same applies to values that are somewhere between observed and missing. So-called coarse data means that we only know that a data point belongs to a certain subset of all possibilities. So-called soft or fuzzy data generalises this further by giving weights to the possibilities. In [10], different ways of handling soft data are studied in context of variational Bayesian learning. A simple example is given in Figure 4.3. The approach called virtual evidence is recommended based on both theory and experimentation with real image data. Also, a justification is given for the standard preprocessing step of adding a tiny amount of noise to the data, when a continuous-valued model is used for discrete-valued data.

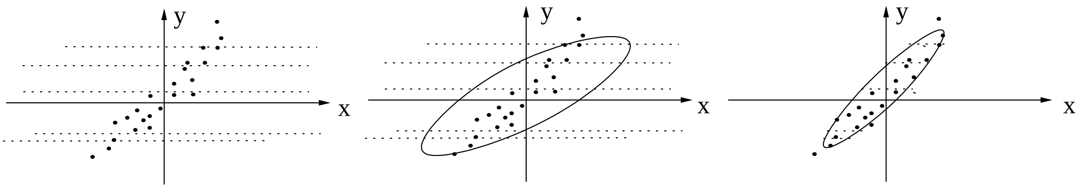


Figure 4.3: Some x-values of the data are observed only partially. They are marked with dotted lines representing their confidence intervals. Left: A simple data set for a factor analysis problem. Middle: In the compared approach, the model needs to adjust to cover the distributions. Right: In the proposed approach, the partially observed values are reconstructed based on the model.

4.3 Building blocks for variational Bayesian learning

In graphical models, there are lots of possibilities to build the model structure that defines the dependencies between the parameters and the data. To be able to manage the variety, we have designed a modular software package using C++/Python called the Bayes Blocks [11]. The theoretical background on which it is based on, was published in [12].

The design principles for Bayes Blocks have been the following. Firstly, we use standardized building blocks that can be connected rather freely and can be learned with local learning rules, i.e. each block only needs to communicate with its neighbors. Secondly, the system should work with very large scale models. We have made the computational complexity linear with respect to the number of data samples and connections in the model.

The building blocks include Gaussian variables, summation, multiplication, and nonlinearity. Recently, several new blocks were implemented including mixture-of-Gaussians and rectified Gaussians [13]. Each of the blocks can be a scalar or a vector. Variational Bayesian learning provides a cost function which can be used for updating the variables as well as optimizing the model structure. The derivation of the cost function and learning rules is automatic which means that the user only needs to define the connections between the blocks.

Examples of structures which can be build using the Bayes Blocks library can be found in Figure 4.4 in the following subsection as well as [12, 14].

Hierarchical modeling of variances

In many models, variances are assumed to have constant values although this assumption is often unrealistic in practice. Joint modeling of means and variances is difficult in many learning approaches, because it can give rise to infinite probability densities. In Bayesian methods where sampling is employed, the difficulties with infinite probability densities are avoided, but these methods are not efficient enough for very large models. Our variational Bayesian method [14], which is based on the building blocks framework, is able to jointly model both variances and means efficiently.

The basic building block in our models is the variance node, which is a time-dependent Gaussian variable $u(t)$ controlling the variance of another time-dependent Gaussian variable $\xi(t)$

$$\xi(t) \sim \mathcal{N}\big(\mu_\xi(t), \exp[-u(t)]\big)$$

Here $\mathcal{N}(\mu, \sigma^2)$ is the Gaussian distribution with mean μ and variance σ^2, and $\mu_\xi(t)$ is the mean of $\xi(t)$ given by other parts of the model.

Figure 4.4 shows three examples of usage of variance nodes. The first model does not have any upper layer model for the variances. There the variance nodes are useful as such for generating super-Gaussian distributions for $\mathbf{s}$, enabling us to find independent components. In the second model the sources can model concurrent changes in both the observations $\mathbf{x}$ and the modeling error of the observations through variance nodes $\mathbf{u}_x$. The third model is a hierarchical extension of the linear ICA model. The correlations and concurrent changes in the variances $\mathbf{u}_s(t)$ of conventional sources $\mathbf{s}(t)$ are modeled by higher-order variance sources $\mathbf{r}(t)$.

We have used the model of Fig. 4.4(c) for finding variance sources from biomedical data containing MEG measurements from a human brain [14]. The signals are contaminated by external artefacts such as digital watch, heart beat, as well as eye movements and blinks. The most prominent feature in the area we used from the dataset is the biting artefact. There the muscle activity contaminates many of the channels starting after 1600 samples.

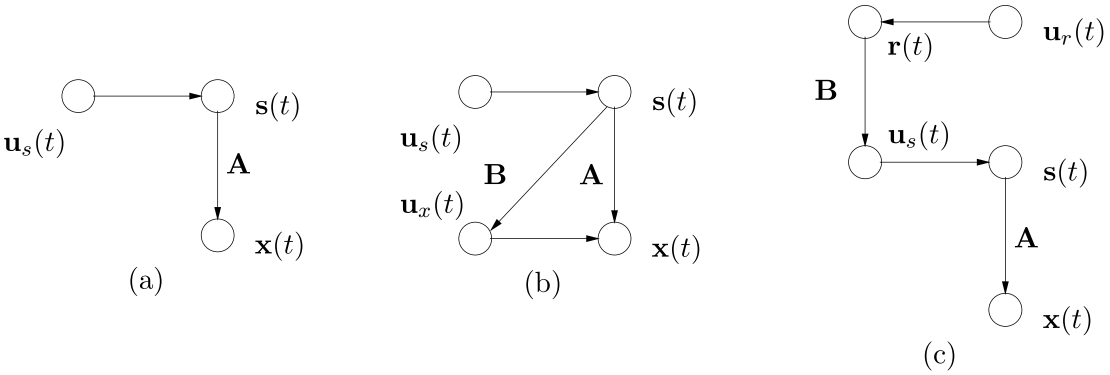


Figure 4.4: Various model structures utilizing variance nodes. Observations are denoted by $\mathbf{x}$, linear mappings by $\mathbf{A}$ and $\mathbf{B}$, sources by $\mathbf{s}$ and $\mathbf{r}$, and variance nodes by $\mathbf{u}$.

Some of the estimated ordinary sources $\mathbf{s}(t)$ are shown in Figure 4.5(a). The variance sources $\mathbf{r}(t)$ that were discovered are shown in Figure 4.5(b). The first variance source clearly models the biting artefact. This variance source integrates information from several conventional sources, and its activity varies very little over time. The second variance source appears to represent increased activity during the onset of the biting, and the third variance source seems to be related to the amount of rhythmic activity on the sources.

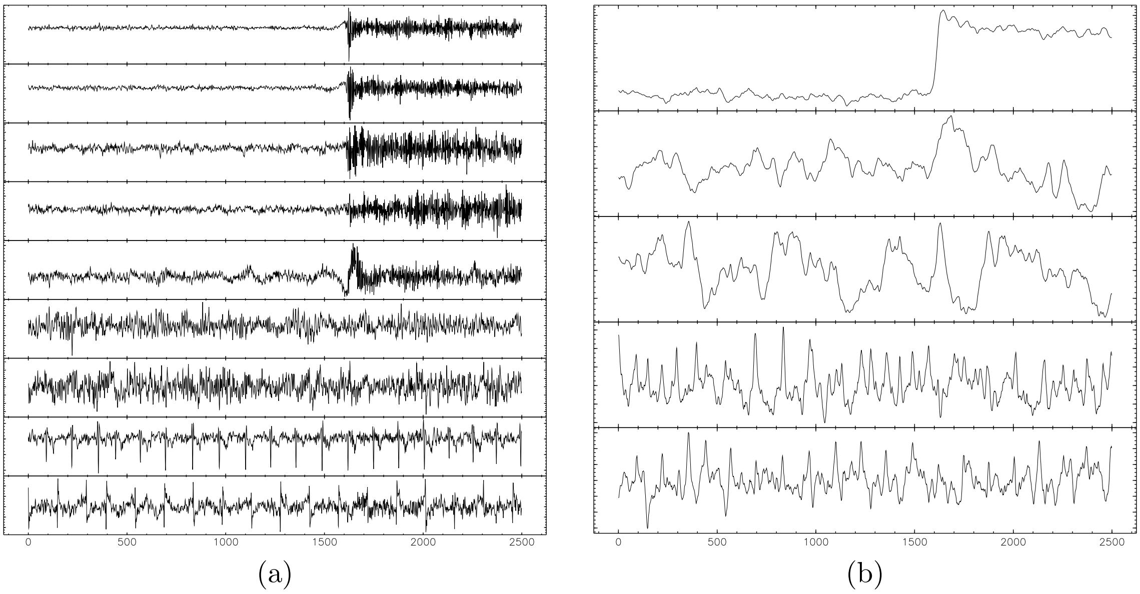


Figure 4.5: (a) Sources $\mathbf{s}(t)$ (nine out of 50) estimated from the data. (d) Variance sources $\mathbf{r}(t)$ which model the regularities found from the variances of the sources [14].

4.4 Nonlinear and non-negative blind source separation

Linear factor analysis (FA) [15] models the data so that it has been generated by sources through a linear mapping with additive noise. Under low noise the method reduces to principal component analysis (PCA). These methods are insensitive to orthogonal rotations of the sources as they only use second order statistics. This can be resolved in the low noise case by independent component analysis (ICA) by assuming independence of the sources and using higher order information [15]. Non-negativity constraints provide an alternative method of resolving the rotation indeterminacy. These methods can be used for blind source separation (BSS) of the sources.

We have applied variational Bayesian learning to nonlinear FA and BSS where the generative mapping from sources to data is not restricted to be linear. The general form of the model is

$$\mathbf{x}(t) = \mathbf{f}(\mathbf{s}(t), \boldsymbol{\theta}_f) + \mathbf{n}(t)\,. \tag{4.3}$$

This can be viewed as a model about how the observations were generated from the sources. The vectors $\mathbf{x}(t)$ are observations at time t, $\mathbf{s}(t)$ are the sources, and $\mathbf{n}(t)$ the noise. The function $\mathbf{f}(\cdot)$ is a mapping from source space to observation space parametrized by $\boldsymbol{\theta}_f$.

BSS and FA in problems with nonlinear mixing

In an earlier work [6] we have used multi-layer perceptron (MLP) network with tanh-nonlinearities to model the mapping $\mathbf{f}$:

$$\mathbf{f}(\mathbf{s}; \mathbf{A}, \mathbf{B}, \mathbf{a}, \mathbf{b}) = \mathbf{B} \tanh(\mathbf{A}\mathbf{s} + \mathbf{a}) + \mathbf{b}\,. \tag{4.4}$$

The mapping $\mathbf{f}$ is thus parameterized by the matrices $\mathbf{A}$ and $\mathbf{B}$ and bias vectors $\mathbf{a}$ and $\mathbf{b}$. MLP networks are well suited for nonlinear FA and BSS. First, they are universal function approximators which means that any type of nonlinearity can be modeled by them in principle. Second, it is easy to model smooth, nearly linear mappings with them. This makes it possible to learn high dimensional nonlinear representations in practice.

The more accurate linearisation presented in Section 4.2 increases stability of the algorithm in cases with a large number of sources when the posterior variances of the last weak sources are typically large.

Using the MLP network in nonlinear BSS leads to an optimisation problem with many local minima. This makes the method sensitive to initialisation. Originally we have used linear PCA to initialise the posterior means of the sources. This can lead to suboptimal results if the mixing is strongly nonlinear. In [16] nonlinear kernel PCA has been used for initialisation. With a proper choice of the kernel, this can lead to significant improvement in separation results.

An alternative hierarchical nonlinear factor analysis (HNFA) model for nonlinear BSS using the building blocks presented in Section 4.3 was introduced in [17]. HNFA is applicable to larger problems than the MLP based method, as the computational complexity is linear with respect to the number of sources. The efficient pruning facilities of Bayes Blocks also allow determining whether the nonlinearity is really needed and pruning it out when the mixing is linear, as demonstrated in [18].

Post-nonlinear FA and BSS

Our work [20] restricts the general nonlinear mapping in (4.3) to the important case of post-nonlinear (PNL) mixtures. The PNL model consists of a linear mixture followed by component-wise nonlinearities acting on each output independently from the others:

$$x_i(t) = f_i\left[\mathbf{a}_i^T\mathbf{s}(t)\right] + n_i(t) \qquad i = 1, \ldots, n \tag{4.5}$$

The vector $\mathbf{a}_i$ in this equation denotes the i:th row of the mixing matrix $\mathbf{A}$. The sources $\mathbf{s}(t)$ are assumed to have Gaussian distributions in our model called *post-nonlinear factor analysis* (PNFA). The sources found with PNFA can be further rotated using any algorithm for linear ICA in order to obtain independent sources.

The development of PNFA was motivated by the comparison experiments [19] where we showed that the existing PNL methods cannot separate globally invertible post-nonlinear mixtures with non-invertible post-nonlinearities. The proposed technique learns the generative model of the observations and therefore it is applicable to such complex PNL mixtures. In [20], we show that PNFA can achive separation of signals in a very challenging BSS problem.

Non-negative BSS by rectified factor analysis

Linear factor models with non-negativity constraints have received a great deal of interest in a number of problem domains. In the variational Bayesian framework, positivity of the factors can be achieved by putting a non-negatively supported prior on the factors. The rectified Gaussian distribution is particularly convenient, as it is conjugate to the Gaussian likelihood arising in the FA model. Unfortunately, this solution has a serious technical limitation: it includes in practice the assumption that the factors have sparse distributions, meaning that the probability mass is concentrated near zero. This is because the location parameter of the prior has to be fixed to zero; otherwise the potentials arising both to the location and to the scale parameter become very awkward.

A way to circumvent the above mentioned problems is to reformulate the model using rectification nonlinearities. This can be expressed in the formalism of Eq. (4.3) using the following nonlinearity

$$\mathbf{f}(\mathbf{s};\, \mathbf{A}) = \mathbf{A}\,\mathbf{cut}(\mathbf{s}) \tag{4.6}$$

where **cut** is the componentwise rectification (or cut) function such that $[\mathbf{cut}(\mathbf{s})]_i = \max(s_i, 0)$. In [21], a variational learning procedure was derived for the proposed model and it was shown that it indeed overcomes the problems that exist with the related approaches. In Section 4.7 an application of the method to the analysis of galaxy spectra is presented.

4.5 Dynamic modelling using nonlinear state-space models

Nonlinear state-space models

In many cases, measurements originate from a dynamical system and form time series. In such cases, it is often useful to model the dynamics in addition to the instantaneous observations. We have extended the nonlinear factor analysis model by adding a nonlinear model for the dynamics of the sources $\mathbf{s}(t)$ [7]. This results in a state-space model where the sources can be interpreted as the internal state of the underlying generative process.

The nonlinear static model of Eq. (4.3) can easily be extended to a dynamic one by adding another nonlinear mapping to model the dynamics. This leads to source model

$$\mathbf{s}(t) = \mathbf{g}(\mathbf{s}(t-1), \boldsymbol{\theta}_g) + \mathbf{m}(t)\,, \tag{4.7}$$

where $\mathbf{s}(t)$ are the sources (states), $\mathbf{m}$ is the Gaussian noise, and $\mathbf{g}(\cdot)$ is a vector containing as its elements the nonlinear functions modelling the dynamics.

As in nonlinear factor analysis, the nonlinear functions are modelled by MLP networks. The mapping $\mathbf{f}$ has the same functional form (4.4). Since the states in dynamical systems are often slowly changing, the MLP network for mapping $\mathbf{g}$ models the change in the value of the source:

$$\mathbf{g}(\mathbf{s}(\mathbf{t-1})) = \mathbf{s}(t-1) + \mathbf{D}\tanh[\mathbf{C}\mathbf{s}(t-1) + \mathbf{c}] + \mathbf{d}\,. \tag{4.8}$$

An important advantage of the proposed method is its ability to learn a high-dimensional latent source space. We have also reasonably solved computational and over-fitting problems which have been major obstacles in developing this kind of unsupervised methods thus far. Potential applications for our method include prediction and process monitoring, control and identification.

Detection of process state changes

One potential application for the nonlinear state-space model is process monitoring. In [22], variational Bayesian learning was shown to be able to learn a model which is capable of detecting an abrupt change in the underlying dynamics of a fairly complex nonlinear process. The process was artificially generated by nonlinearly mixing some of the states of three independent dynamical systems: two independent Lorenz processes and one harmonic oscillator. The nonlinear dynamic model was first estimated off-line using 1000 samples of the observed process. The model was then fixed and applied on-line to new observations with artificially generated changes of the dynamics.

Figures 4.6 and 4.7 show an experiment with a change generated at time instant T_{ch}, when the underlying dynamics of one of the Lorenz processes abruptly changes. The change detection method based on the estimated model readily detects the change raising an alarm after the time of change. The method is also able to find out in which states the change occurred (see Fig. 4.7) as the reason for the detected change can be localised by analysing the structure of the cost function.

Stochastic nonlinear model-predictive control

For being able to control the dynamical system, control inputs are added to the nonlinear state-space model. In [23], we study three different control schemes in this setting. Direct control is based on using the internal forward model directly. It is fast to use, but requires the learning of a policy mapping, which is hard to do well. Optimistic inference control is a novel method based on Bayesian inference answering the question: “Assuming success

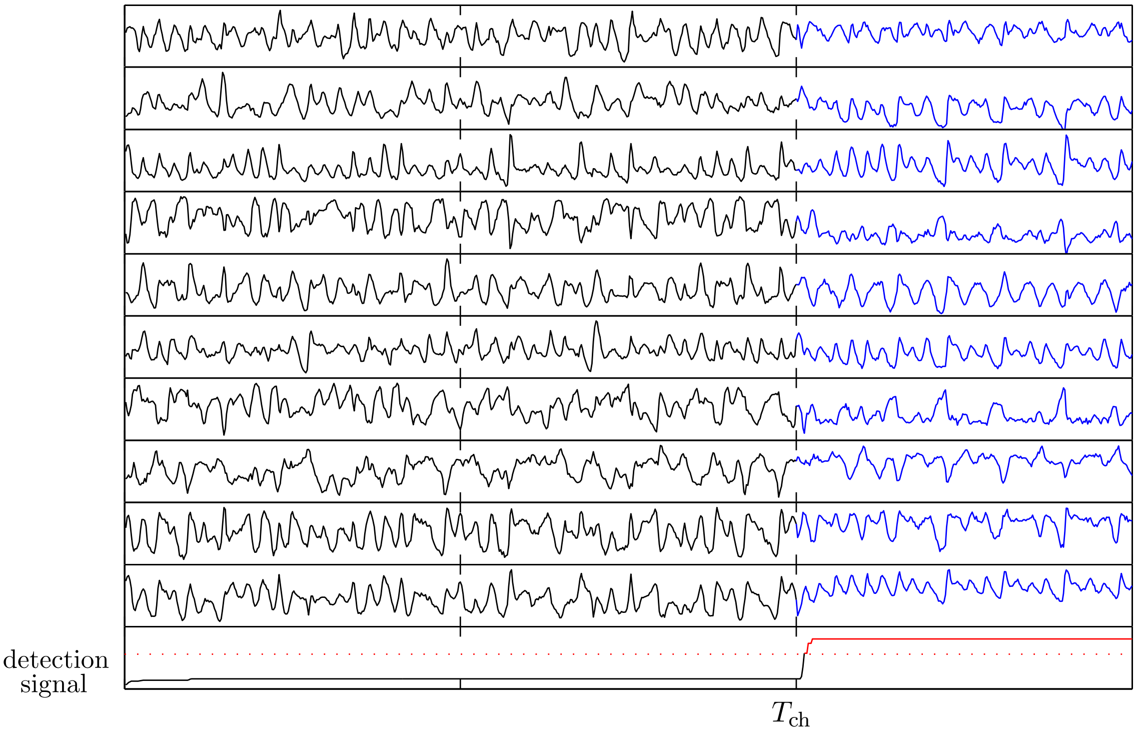


Figure 4.6: The monitored process (10 time series above) with the change simulated at T_{ch}. The change has been detected using the estimated model, the alarm signal is shown below.

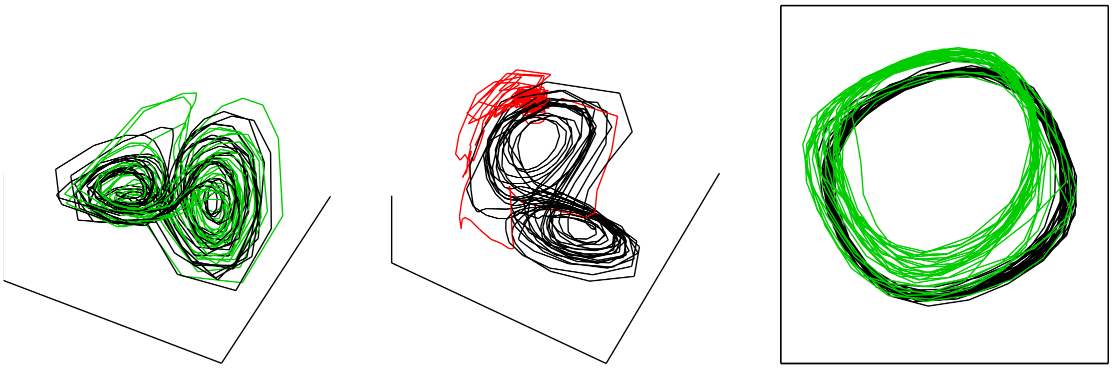


Figure 4.7: The estimated process states reconstructing the two original Lorenz processes and harmonic oscillator. The values after T_{ch} are shown as coloured curves. The cost contribution of the second process drastically changes after the time of change, which is used to localise the reason of the change.

in the end, what will happen in near future?" It is based on a single probabilistic inference but unfortunately neither of the two tested inference algorithms works well with it. The third control scheme is stochastic nonlinear model-predictive control, which is based on optimising control signals based on maximising a utility function.

Figure 4.8 shows simulations with a cart-pole swing-up task. The results confirm that selecting actions based on a state-space model instead of the observation directly has many benefits: First, it is more resistant to noise because it implicitly involves filtering. Second,

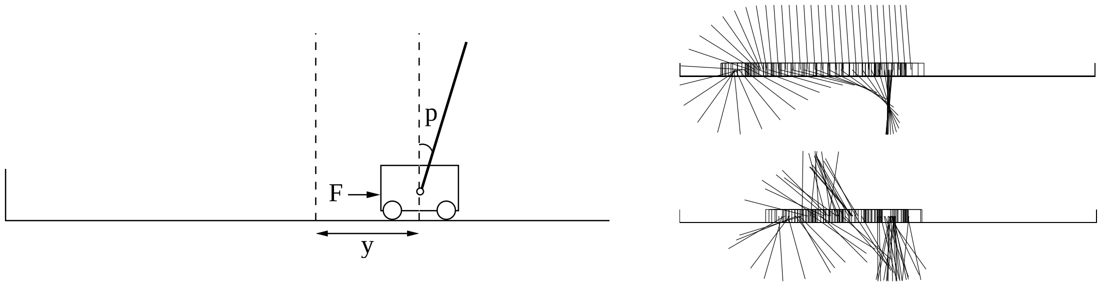


Figure 4.8: Left: the cart-pole system. The goal is to swing the pole to an upward position and stabilise it without hitting the walls. The cart can be controlled by applying a force to it. Top left: the pole is successfully swinged up by moving first to the left and then right. Bottom right: our controller works quite reliably even in the presence of serious observation noise [23].

the observations (without history) do not always carry enough information about the system state. Third, when nonlinear dynamics are modelled by a function approximator such as an multilayer perceptron network, a state-space model can find such a representation of the state that it is more suitable for the approximation and thus more predictable [23].

4.6 Relational models

Formerly, we have divided our models into two categories: static and dynamic. In static modelling, each observation or data sample is independent of the others. In dynamic models, the dependencies between consecutive observations are modelled. The generalisation of both is that the relations are described in the data itself, that is, each observation might have a different structure.

Many models have been developed for relational discrete data, and for data with nonlinear dependencies between continuous values. In [24], we combine two of these methods, relational Markov networks and hierarchical nonlinear factor analysis, resulting in using nonlinear models in a structure determined by the relations in the data. Experimental setup in the board game Go is depicted in Figure 4.9. The task is the collective regression of survival probabilities of blocks. The results suggest that regression accuracy can be improved by taking into account both relations and nonlinearities.

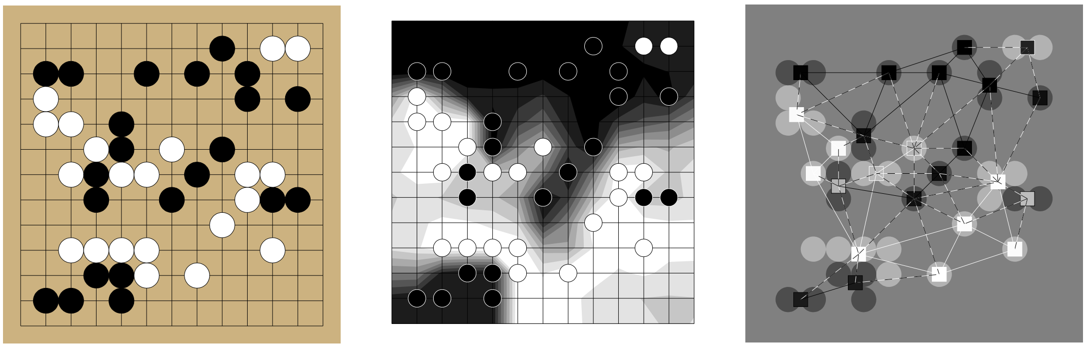


Figure 4.9: The leftmost subfigure shows the board of a Go game in progress. In the middle, the expected owner of each point is visualised with the shade of grey. For instance, the two white stones in the upper right corner are very likely to be captured. The rightmost subfigure shows the blocks with their expected owner as the colour of the square. Pairs of related blocks are connected with a line which is dashed when the blocks are of opposing colours.

Many real world sequences such as protein secondary structures or shell logs exhibit rich internal structures. Logical hidden Markov models have been proposed as one solution. They deal with logical sequences, i.e., sequences over an alphabet of logical atoms. This comes at the expense of a more complex model selection problem. Indeed, different abstraction levels have to be explored. In [25], we propose a novel method for selecting logical hidden Markov models from data called SAGEM. SAGEM combines generalized expectation maximization, which optimizes parameters, with structure search for model selection using inductive logic programming refinement operators. We provide convergence and experimental results that show SAGEM's effectiveness.

4.7 Applications to astronomy

We have applied rectified factor analysis [21] described in Section 4.4 to the analysis of real stellar population spectra of elliptical galaxies. Ellipticals are the oldest galactic systems in the local universe and are well studied in physics. The hypothesis that some of these old galactic systems may actually contain young components is relatively new. Hence, we have investigated whether a set of stellar population spectra can be decomposed and explained in terms of a small set of unobserved spectral prototypes in a data driven but yet physically meaningful manner. The positivity constraint is important in this modelling application, as negative values of flux would not be physically interpretable.

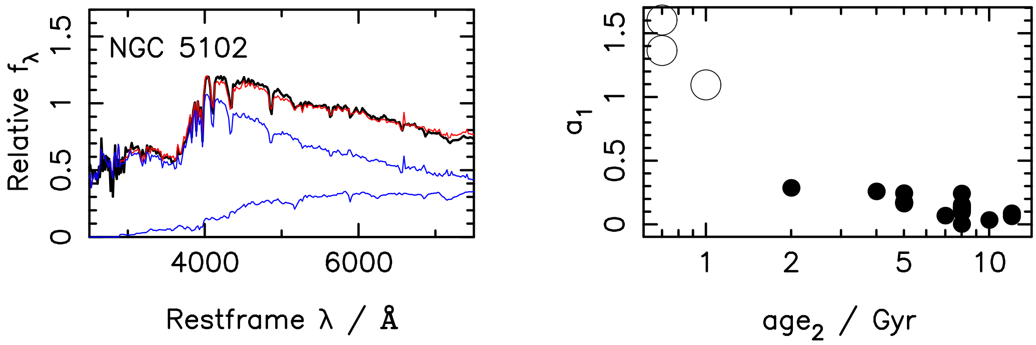


Figure 4.10: Left: the spectrum of a galaxy with its decomposition to a young and old component. Right: the age of the dominating stellar population against the mixing coefficient of the young component.

Using a set of 21 real stellar population spectra, we found that they can indeed be decomposed to prototypical spectra, especially to a young and old component [26]. Figure 4.10 shows one spectrum and its decomposition to these two components. The right subfigure shows the ages of the galaxies, known from a detailed astrophysical analysis, plotted against the first weight of the mixing matrix. The plot clearly shows that the first component corresponds to a galaxy containing a significant young stellar population.

References

[1] D. J. C. MacKay, *Information Theory, Inference, and Learning Algorithms.* Cambridge University Press, 2003.

[2] H. Lappalainen and J. Miskin. Ensemble learning. In M. Girolami, editor, *Advances in Independent Component Analysis*, Springer, 2000, pages 75–92.

[3] M. I. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul. An introduction to variational methods for graphical models. In M. I. Jordan, editor, *Learning in Graphical Models* MIT Press, 1999, pages 105–161.

[4] A. Honkela and H. Valpola. Variational learning and bits-back coding: an information-theoretic view to Bayesian learning. *IEEE Transactions on Neural Networks*, 15(4):267–282, 2004.

[5] A. Ilin, and H. Valpola. On the effect of the form of the posterior approximation in variational learning of ICA models. *Neural Processing Letters*, 22(2):183–204, 2005.

[6] H. Lappalainen and A. Honkela. Bayesian nonlinear independent component analysis by multi-layer perceptrons. In Mark Girolami, editor, *Advances in Independent Component Analysis*, pages 93–121. Springer-Verlag, Berlin, 2000.

[7] H. Valpola and J. Karhunen. An unsupervised ensemble learning method for nonlinear dynamic state-space models. *Neural Computation*, 14(11):2647–2692, 2002.

[8] A. Honkela. Approximating nonlinear transformations of probability distributions for nonlinear independent component analysis. In *Proc. 2004 IEEE Int. Joint Conf. on Neural Networks (IJCNN 2004)*, pages 2169–2174, Budapest, Hungary, 2004.

[9] A. Honkela and H. Valpola. Unsupervised variational Bayesian learning of nonlinear models. In L. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 593–600. The MIT Press, Cambridge, MA, USA, 2005.

[10] T. Raiko. Partially observed values. In *Proc. 2004 IEEE Int. Joint Conf. on Neural Networks (IJCNN 2004)*, pages 2825–2830, Budapest, Hungary, 2004.

[11] H. Valpola, A. Honkela, M. Harva, A. Ilin, T. Raiko, and T. Östman. Bayes Blocks software library. `http://www.cis.hut.fi/projects/bayes/software/`, 2003.

[12] H. Valpola, T. Raiko, and J. Karhunen. Building blocks for hierarchical latent variable models. In *Proc. 3rd Int. Workshop on Independent Component Analysis and Signal Separation (ICA2001)*, pages 710–715, San Diego, California, December 2001.

[13] M. Harva, T. Raiko, A. Honkela, H. Valpola, and J. Karhunen. Bayes Blocks: An implementation of the variational Bayesian building blocks framework. In *Proc. 21st Conference on Uncertainty in Artificial Intelligence*, pages 259–266. Edinburgh, Scotland, 2005.

[14] H. Valpola, M. Harva, and J. Karhunen. Hierarchical models of variance sources. *Signal Processing*, 84(2):267–282, 2004.

[15] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis.* Wiley, 2001.

[16] A. Honkela, S. Harmeling, L. Lundqvist and H. Valpola. Using kernel PCA for initialisation of variational Bayesian nonlinear blind source separation method. In *Proc. 5th Int. Conf. on Independent Component Analysis and Blind Signal Separation (ICA 2004)*, Vol. 3195 of *Lecture Notes in Computer Science*, Springer-Verlag, pages 790–797, 2004.

[17] H. Valpola, T. Östman, and J. Karhunen. Nonlinear independent factor analysis by hierarchical models. In *Proc. 4th Int. Symp. on Independent Component Analysis and Blind Signal Separation (ICA2003)*, pages 257–262, Nara, Japan, 2003.

[18] A. Honkela, T. Östman, and R. Vigário. Empirical evidence of the linear nature of magnetoencephalograms. In *Proc. 13th European Symp. on Artificial Neural Networks (ESANN 2005)*, pages 285–290, Bruges, Belgium, 2005.

[19] A. Ilin, S. Achard, and C. Jutten. Bayesian versus constrained structure approaches for source separation in post-nonlinear mixtures. In *Proc. of the Int. Joint Conf. on Neural Networks (IJCNN 2004)*, pages 2181–2186, Budapest, Hungary, 2004.

[20] A. Ilin and A. Honkela. Post-nonlinear independent component analysis by variational Bayesian learning. In *Proc. 5th Int. Conf. on Independent Component Analysis and Blind Signal Separation (ICA 2004)*, Vol. 3195 of *Lecture Notes in Computer Science*, Springer-Verlag, pages 766–773, 2004.

[21] M. Harva and A. Kabán. A variational Bayesian method for rectified factor analysis. In *Proc. Int. Joint Conf. on Neural Networks (IJCNN'05)*, pages 185–190. Montreal, Canada, 2005.

[22] A. Ilin, H. Valpola, and E. Oja. Nonlinear dynamical factor analysis for state change detection. *IEEE Transaction on Neural Networks*, 15(3):559–575, 2004.

[23] T. Raiko and M. Tornio. Learning nonlinear state-space models for control. In *Proc. Int. Joint Conf. on Neural Networks (IJCNN'05)*, pages 815–820, Montreal, Canada, 2005.

[24] T. Raiko. Nonlinear relational Markov networks with an application to the game of Go. In Proc. Int. Conf. on Artificial Neural Networks (ICANN 2005), pages 989–996, Warsaw, Poland, September 2005.

[25] K. Kersting and T. Raiko. 'Say EM' for selecting probabilistic models for logical sequences. In *Proc. 21st Conference on Uncertainty in Artificial Intelligence*, pages 300–307, Edinburgh, Scotland, July 2005.

[26] L. Nolan, M. Harva, A. Kabán, and S. Raychaudhury. A data-driven Bayesian approach for finding young stellar populations in early-type galaxies from their UV-optical spectra. *Monthly Notices of the Royal Astronomical Society*, 366(1):321–338, 2006.

Chapter 5

Bioinformatics

Samuel Kaski, Janne Nikkilä, Merja Oja, Leo Lahti, Jarkko Venna, Eerika Savia, Arto Klami

5.1 Introduction

New so-called high-throughput measurement techniques have made possible genome-wide studies of gene function. Gene expression, gene regulation, protein content, protein interaction, and metabolic profiles can be measured and combined with the genetic sequence. The current challenge is to extract meaningful findings from the noisy and incomplete data masses, collected into both community resource and private data banks. The data needs to be analyzed, mined, understood, and taken into account in further experiments, which makes data analysis an integral part of biomedical research. Successful genome-wide analyses would allow a completely novel systems-level view into a biological organism.

Combining the different kinds of data produces new systems-level hypotheses about gene function and regulation, and ultimately functioning of biological organisms. We develop probabilistic modeling and statistical data analysis methods to advance this field.

The project is carried out in collaboration with experts of the biomedical areas and with the other bioinformatics group of the laboratory that belongs to the From Data to Knowledge research unit.

5.2 Yeast systems biology

A major component of systems biology is integration of information from multiple sources. The integration is far from trivial since the data types and scales can vary dramatically. We propose a new framework for systems biology that enables focusing on relevant variation in the data sets, the relevance being determined by other, auxiliary data sets. Dependency modeling and learning metrics methods (Section 6) provide a state of the art tool for this. Together with Docent Christophe Roos from Medicel Ltd. we have developed and applied them for yeast systems biology, especially for exploring yeast stress reaction and its regulation by transcription factor proteins.

Defining yeast stress reaction

Yeast is a key model organism in biological research and process industry. The majority of experiments and utilization of yeast take place by modifying yeast's environmental conditions. A problem is that this always shifts the yeast's biological state from the optimal to more or less stressful state that affects yeast's behaviour. Understanding the yeast stress is thus of crucial importance.

It is believed that yeast as a unicellular organism has a special set of genes that are always activated under any environmental stress. Still, it is very difficult to define an explicit, parametric model for the stress reaction. We have used a novel method to define the stress behavior of the yeast in a data driven way: stress reaction are the effects that are common across different stress treatments.

We have applied a version of gCCA (Section 6) in a novel way for data-driven extraction of the stress effect from the multiple gene expression data sets measured under various stress treatments [5, 6]. The use of gCCA for feature extraction produces a lower dimensional subspace of the original joint data space of all the data sets. That subspace maximally preserves the dependencies between the original data sets. Figure 5.1 presents an overview of the application for gCCA to yeast stress extraction.

In [3] it was demonstrated that even better representation can be achieved by using a non-parametric measure for the dependency in place of the correlation used in gCCA, but with an increased computational cost.

Exploring regulation of yeast gene expression

The biological state of the cell is for a large part defined by which genes are expressed at a certain moment. The regulation of gene expression is thus the key for understanding, for example, the reasons why some cells are transformed to cancer cells. The regulation of expression has been under intensive study during past five years, but it has proved to be extremely difficult to model with detailed statistical models because of the small sample sizes. We search for the hints which genes are dependent on regulating proteins by general purpose methods that work in a data-driven way and utilize the latent group structure of the genes by clustering.

Gene expression is largely regulated by a set proteins called *transcription factors*. They affect gene expression by binding a gene's promoter region, and their type and configuration on the promoter determines in part the activity of the gene. We explored yeast gene regulatory mechanisms with *associative clustering (AC)* (Section 6), by searching for gene groups that are maximally dependent by expression and by transcription factor binding (see Figure 5.2) [4]. We found statistically significant dependency, confirmed the results with known regulatory mechanisms, and generated hypotheses for new regulatory interactions for many expression data sets, including the expression under environmental stress [1, 7, 6].

Additionally, *discriminative clustering (DC)* (Section 6) was applied to explore the regulation of yeast stress genes [2]. Stress genes should behave similarly in all stress treaments, and potentially be regulated by certain regulators (MSN2/4). We clustered yeast gene expression profiles measured in stress treatments, and supervised the clustering by the change of the behavior after the potential regulators were knocked out. This focused the clusters on gene expression regulated by MSN2/4.

We identified a subset of genes that are upregulated in all stress conditions, but only when regulators MSN2 and MSN4 are functional. Stress genes found in an independent study were strongly enriched in the discovered subset.

References

[1] Samuel Kaski, Janne Nikkilä, Janne Sinkkonen, Leo Lahti, Juha Knuuttila, and Christophe Roos. Associative clustering for exploring dependencies between functional genomics data sets. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, Special Issue on Machine Learning for Bioinformatics – Part 2*, 2(3):203–216, 2005.

[2] Samuel Kaski, Janne Nikkilä, Eerika Savia, and Christophe Roos. Discriminative clustering of yeast stress response. In Udo Seiffert, Lakhmi Jain, and Patric Schweizer,

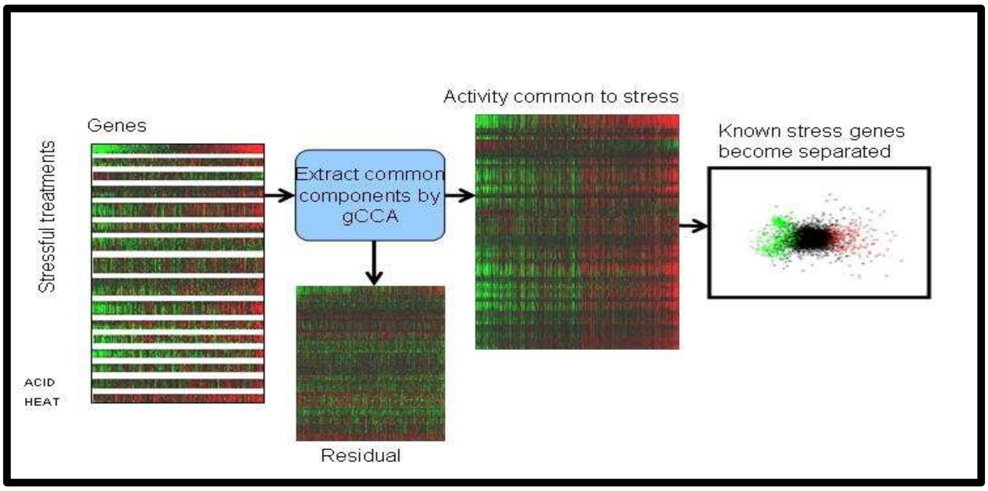

Stressful treatments
Genes
ACID
HEAT
Extract common components by gCCA
Activity common to stress
Residual
Known stress genes become separated

Figure 5.1: Dimensionality reduction by gCCA extracts common properties of data sets. On the left are the original expression data sets from various stress treatments. Red denotes the genes upregulated during stress and, respectively, green the genes that are downregulated. Applying gCCA to the concatenated data sets in a novel way finds a linear subspace that maximally preserves the dependencies between the original variables (in the middle). In the rightmost figure, when the data is projected on the two first components, the known stress genes (red and green dots) become separated from the rest (black dots).

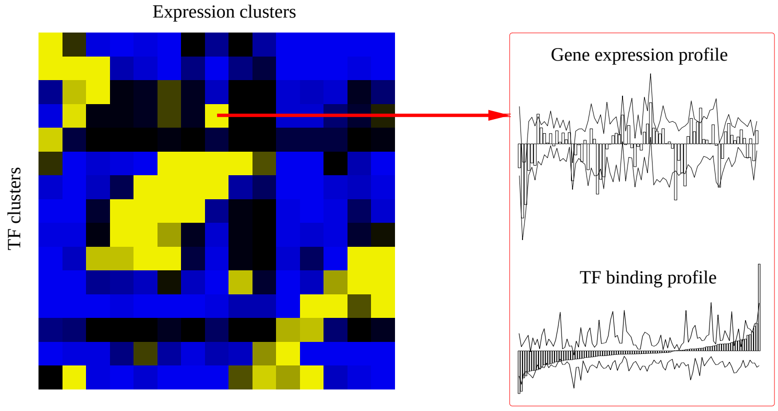


Figure 5.2: Example of a significant association between transcription factor binding and gene expression patterns identified by AC. The yellow cross cluster is associated with cell cycle, and reveals both known and novel dependencies between gene expression and TF binding. The upper profile shows the average expression profile (bars) of the cluster, and the confidence intervals (curves). The lower figure show the average TF-binding profile of the clusters with confidence intervals. In this cluster there were two reliable TF bindings (the rightmost bars in the lower figure), SIP4 and SFL1, of which SIP4 could be verified from the literature, and SFL1 is a new potential regulator for the genes in this cluster.

editors, *Bioinformatics using Computational Intelligence Paradigms*, pages 75–92. Springer,Berlin, 2005.

[3] Arto Klami and Samuel Kaski. Non-parametric dependent components. In *Proceedings of ICASSP'05, IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages V–209–V–212. IEEE, 2005.

[4] T.I. Lee, N.J. Rinaldi, F. Robert, D.T. Odom, Z. Bar-Joseph, G.K. Gerber, N.M. Hannett, C.T. Harbison, C.M. Tomphson, I. Simon, J. Zeitlinger, E.G. Jennings, H.L. Murray, D.B. Gordon, B. Ren, J.J. Wyrick, J.-B. Tagne, T.L. Volkert, E.Fraenkel, D.K Gifford, and R.A. Young. Transcriptional regulatory networks in saccharomyces cerevisiae. *Science*, 298:799–804, 2002.

[5] J. Nikkilä, C. Roos, and S. Kaski. Exploring dependencies between yeast stress genes and their regulators. In Zheng Rong Yang, Richard Everson, and Hujun Yin, editors, *Proceedings of International Conference on Intelligent Data Engineering and Automated Learning (IDEAL 2004)*, pages 92–98. Springer, 2004.

[6] Janne Nikkilä, Christophe Roos, Eerika Savia, and Samuel Kaski. Explorative modeling of yeast stress response and its regulation with gCCA and associative clustering. *International Journal of Neural Systems*, 15(4):237–246, 2005.

[7] Janne Sinkkonen, Janne Nikkilä, Leo Lahti, and Samuel Kaski. Associative clustering. In Boulicaut, Esposito, Giannotti, and Pedreschi, editors, *Machine Learning: ECML2004 (Proceedings of the ECML'04, 15th European Conference on Machine Learning)*, pages 396–406. Springer, Berlin, 2004.

5.3 Comparative functional genomics

Comparative genomics studies the similarity of the genes in different species. This is of utmost importance when animal models are used to study human diseases and the inferences based on an animal models are translated to human. The basis of this translation are usually the *orthologous genes*.

Determination of orthologous genes, between organisms, relies on similarities in their DNA sequence. However, this does not guarantee their functional similarity. We explored the *functional* dependencies between human and mouse orthologous gene expression by clustering them with associative clustering [1, 2]. In collaboration with the group of Eero Castrén from Neuroscience Center, University of Helsinki, we confirmed the expected functional similarity for many orthologous genes, but we also found some unexpected differences. These differences suggest deviations of genes' functions during evolution and show evidence in favour of using both functional and sequence information when determining homologous genes.

References

[1] Samuel Kaski, Janne Nikkilä, Janne Sinkkonen, Leo Lahti, Juha Knuuttila, and Christophe Roos. Associative clustering for exploring dependencies between functional genomics data sets. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, Special Issue on Machine Learning for Bioinformatics – Part 2*, 2(3):203–216, 2005.

[2] Janne Sinkkonen, Janne Nikkilä, Leo Lahti, and Samuel Kaski. Associative clustering. In Boulicaut, Esposito, Giannotti, and Pedreschi, editors, *Machine Learning: ECML2004 (Proceedings of the ECML'04, 15th European Conference on Machine Learning)*, pages 396–406. Springer, Berlin, 2004.

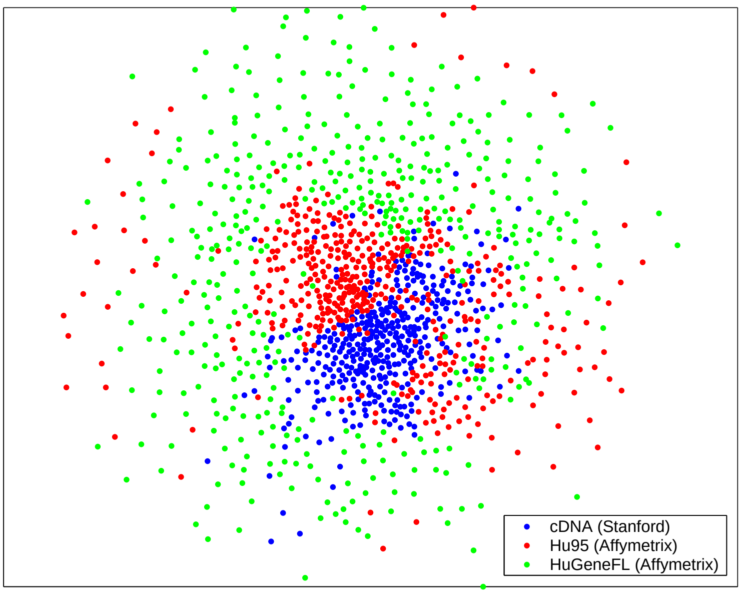


Figure 5.3: Sample visualization of the gene expression atlas by curvilinear component analysis. Each dot denotes one microarray; the colors show the measurement platform.

5.4 Gene expression atlas

A large community-resource or private gene expression databank consists of numerous data sets submitted by several parties. They may have been measured for different purposes, with different treatments and methods in different laboratories. Several such databanks have been established and they continue to grow. A key challenge is how to best use the databanks to support further research. Currently information in these databanks is accessed using queries on the imperfect meta-data, that is, textual annotations and descriptions. In the future more sophisticated search methods, that take the actual data into account, are needed. Our first study [1] aims at creating a visual interface that reveals similarities of data sets.

We compared several different visualization methods in the task of visualizing a large collection of gene expression arrays. A nonlinear dimensionality reduction algorithm, curvilinear component analysis, had the best performance of the methods compared. We also verified that the main sources of variance in the data were also evident from the visualization. In this case the data turned out to be mostly organized based on the type of platform used to perform the experiment. Thus the simple preprocessing method used was not able to make the data commensurable. This is also very evident in Fig. 5.3; the experiments done on each platform create coherent groups and there is only a small amount of mixing of arrays from different platforms. Thus, we were able to produce a visualization organized according to the main sources of variation in the data, and when better preprocessing methods are developed, it will be possible to produce a useful visual interface to a gene expression data bank.

References

[1] Jarkko Venna and Samuel Kaski. Visualized atlas of a gene expression databank. In *Proceedings of Symposium of Knowledge Representation in Bioinformatics*, pages 30–36, Espoo, Finland, 2005.

5.5 Genomics of human endogenous retroviruses

About eight per cent of human DNA consists of remains of specific kinds of transposons[1], called *human endogenous retroviruses (HERV)*. Human retroviruses, such as HIV, in general are viruses capable of copying their genetic code to the DNA of humans, and they become endogenous once they have been copied to the germ-line. Human endogenous retroviruses are remains from ancient infections.

Human endogenous retroviruses, in contrast to some other human transposons, are not capable of moving any longer but it has been suggested that they may have functions in regulating the activity of human genes, and may produce proteins under some conditions [1].

One of the first steps in understanding HERV function is to classify HERVs into families. We have studied the relationships of existing HERV families and tried to detect potentially new HERV families in co-operation with the group of Professor Blomberg, University of Uppsala, [3, 4]. A Median Self-Organizing Map (SOM) [2], a SOM for non-vectorial data, was used to group and visualize a collection of 3661 HERV protein sequences.

The SOM-based analysis was complemented with estimates of the reliability of the results [4]. A novel trustworthiness visualization method was used to estimate which parts of the SOM visualization are reliable and which not. The reliability of extracted interesting HERV groups was verified by a bootstrap procedure suitable for SOM visualization-based analysis. The SOM detected a completely new group of epsilonretroviral sequences and was able to shed light into the relationships of three pre-existing HERV families. The SOM detected a group of ERV9, HERVW, and HUERSP3 sequences which suggested that ERV9 and HERVW sequences may have a common origin.

References

[1] David J. Griffiths. Endogenous retroviruses in the human genome sequence. *Genome Biology*, 2:1017.1–1017.5, 2001.

[2] Teuvo Kohonen and Panu Somervuo. How to make large self-organizing maps for nonvectorial data. *Neural Networks*, 15:945–52, 2002.

[3] Merja Oja, Göran Sperber, Jonas Blomberg, and Samuel Kaski. Grouping and visualizing human endogenous retroviruses by bootstrapping median self-organizing maps. In *CIBCB 2004. IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, 7-8 October, San Diego, USA.*, 2004.

[4] Merja Oja, Göran O. Sperber, Jonas Blomberg, and Samuel Kaski. Self-organizing map-based discovery and visualization of human endogenous retroviral sequence groups. *International Journal of Neural Systems*, 15(3):163–179, 2005.

[1] parts of genome capable of moving or copying themselves in the genome

Chapter 6

Dependency exploration and learning metrics

Samuel Kaski, Jaakko Peltonen, Janne Nikkilä, Jarkko Venna, Jarkko Salojärvi, Arto Klami, Leo Lahti, Janne Sinkkonen

6.1 Introduction

We develop statistical machine learning methods for extracting useful regularities from large, high-dimensional data sets. We have divided this task of *statistical data mining* or exploration into three subtypes of problems:

- *Unsupervised mining*, where regularities or dependencies *within* one data set are sought. Standard clustering, component models, and data mining belongs to this category. We have recently developed new nonlinear projection methods for information visualization.

- *Supervised mining*, where one data set supervises the mining of another. We have introduced a principle of learning metrics, where the distance metric is learned in a supervised way, to guide subsequent unsupervised learning. Hence, this line of work has alternatively been coined *supervised unsupervised learning.*

- *Dependency mining* or exploration, where the supervision is symmetric and the task is to find dependencies *between* data sets. In this subtask we have introduced new clustering and component models.

Unsupervised mining is the most common task of these three but also the most difficult, because it suffers from the garbage in—garbage out problem; unsupervised learning cannot distinguish relevant from irrelevant variation in a data set. Supervised mining and dependency mining are two solutions to this problem that are applicable in different kinds of settings. In supervised mining, variation within one of the sets is assumed irrelevant and the other set is assumed relevant enough to be useful for supervising the other. An example is analysis of measurement data of cancer tissues where known cancer labels (the other set) are clearly extremely relevant, and variation in tissue samples not related to cancer classes is irrelevant.

In dependency mining the within-set variation is assumed irrelevant in all data sets, and only between-set variation is important. An example is measurement of several noisy signals from a common source, when characteristics of the noise are not known. More examples are given in Section 5.

6.2 Supervised unsupervised learning

Many unsupervised methods rely on a distance measure that tells how far apart two data samples are. Usually the measure is a simple one such as Euclidean distance. However, such measures do not take into account that two samples can differ in many ways, and not all differences are relevant for the analysis. How relevant a difference is depends on what the analyst is interested in; the relevance can vary in different parts of the data space.

In supervised settings we can learn what is relevant by learning a distance measure, that is, a metric. The idea of learning such metrics has been coined the *learning metrics principle* [1].

We assume there is paired data: the primary data $\mathbf{x}$ that we want to explore are paired to *auxiliary data* c that guide the exploration. The learning metrics principle assumes that variation of the primary data is important only to the extent it causes variation in auxiliary data.

Technically, we use an information-geometric definition: the distance d between two close-by data points $\mathbf{x}$ and $\mathbf{x} + d\mathbf{x}$ is defined as the difference between the corresponding distributions of c, measured by the Kullback-Leibler divergence D_{KL}, i.e.,

$$d_L^2(\mathbf{x}, \mathbf{x} + d\mathbf{x}) \equiv D_{\mathrm{KL}}(p(c|\mathbf{x}) \| p(c|\mathbf{x} + d\mathbf{x})) = \; d\mathbf{x}^T \mathbf{J}(\mathbf{x}) d\mathbf{x} \;, \tag{6.1}$$

where $\mathbf{J}(\mathbf{x})$ is the Fisher information matrix. The metric is learned from the data that defines the distributions $p(c|\mathbf{x})$. This yields a Riemannian metric that preserves the topology of the feature space, yet can flexibly change what is locally relevant. The preservation of topology and the capability to ignore noisy dimensions not related to the auxiliary data are demonstrated (with empirical experiments) in [2].

In practice, the learning metrics principle can be applied in two ways. One can estimate $p(c|\mathbf{x})$ first and then plug the new metric, computed from the estimates, into a standard unsupervised method. Another possibility is to more directly insert the new metric into the cost function of a suitable method. Examples of both are given below.

Visualization and component models

The self-organizing map learns to visualize multidimensional data on a two-dimensional regular grid by finding and optimizing winner units for data samples. The winner is the closest map unit for a particular sample, and we simply choose the winner by distance in the learning metric. A practical approximation to the distance that in principle is defined by path integrals is the so-called T-point distance approximation [2] defined as

$$d_T(\mathbf{x}_1, \mathbf{x}_2) = \frac{1}{T} \sum_{i=0}^{T-1} \left(\mathbf{r}^T \mathbf{J} \left(\mathbf{x} + \frac{i}{T} \mathbf{r} \right) \mathbf{r} \right)^{1/2} \tag{6.2}$$

where $\mathbf{r} = \mathbf{x}_2 - \mathbf{x}_1$. This approximation assumes the shortest path is a line and computes the local metric at T points between the start and end point.

The SOM in learning metric has been shown shown to outperform SOM in Euclidean metric, as well as a supervised variant of SOM [2]. It provides visualizations that tell more about the auxiliary data, and on top of that allows visualizing how relevant the input features locally are. Example visualizations of SOMs computed using the learning metric are provided in Figure 6.1.

Metric multidimensional scaling methods (MDS) are used for visualizing similarities of data samples based on a pairwise distance matrix. They construct a low-dimensional

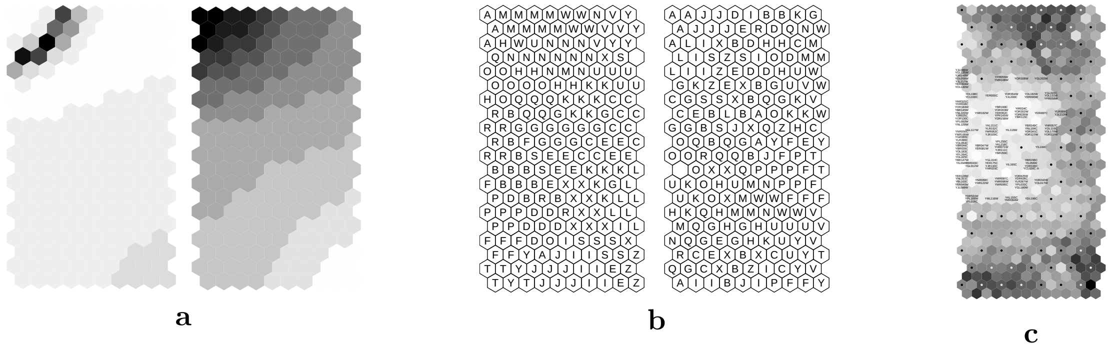


Figure 6.1: Example applications of SOM in learning metrics. **a** Analysis of financial data reveals that the importance of profitability (left) in avoiding bankruptcies depends on the state of the company, here illustrated by the actual profitability indicator (right). **b** Similar written characters are grouped more tighly in learning metrics (left) than in Euclidean metric (right). **c** Part of yeast gene expression data visualization emphasizing the functional categories.

representation for the data that aims to preserve the distance matrix, and are ideal candidates for applying computationally more intensive approximations of learning metrics distances since the distances need to be computed only once.

In this kind of application the assumption of minimal path being a line can be relaxed. Instead, a graph whose edge weights are pairwise T-point distances between data points is formed and a graph search for the minimal path is performed, providing a piece-wise linear path. In [2] this approximation generalized a specific MDS method called Sammon's mapping.

Relevant Component Analysis is a new data analysis tool that finds components of data in the learning metrics. Instead of unsupervised components like principal components, we wish to find components that contain information about, or are relevant for, classes of the data. Such components can be used to reduce dimensionality, in order to explore and visualize class separation, and to study the contribution of original data variables to it. We call the task of finding the components Relevant Component Analysis (RCA).

A classical method for this task is Linear Discriminant Analysis (LDA), which finds linear components but makes restrictive assumptions about the data distribution. We have introduced an improved method that removes the assumptions and finds components by optimizing a simple nonparametric generative model for class labels.

Technically, the method maximizes a simple likelihood criterion:

$$\sum_{(\mathbf{x},c)} \log \hat{p}(c|\mathbf{W}^T\mathbf{x}) \tag{6.3}$$

where the $\mathbf{x}$ are samples, c are their classes, $\hat{p}$ is the nonparametric estimator, and the columns of $\mathbf{W}$ are the component directions. This is equivalent (asymptotically, when $\hat{p}$ is consistent) to maximizing the mutual information of the component projections and the class labels. The components can be interpreted (asymptotically and approximately) as principal components in learning metrics.

The main merits of the new method, compared to other generalizations of LDA, are its theoretical simplicity and good performance. The method has been applied to exploration of sound samples from different phonemes [3], expressions of genes from different functional

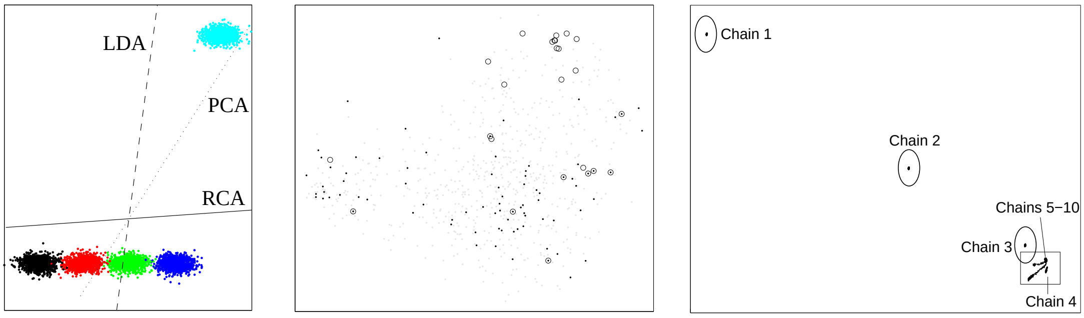


Figure 6.2: Left: The new method, RCA, finds a component that is more informative of classes than the classical methods. Center: two components of gene expression data that are informative of functional classes (only two classes shown). Right: two components of MCMC samples that show differences between sampling chains.

classes [4], and posterior samples from different MCMC chains (see below). Figure 6.2 shows examples on both toy data and real applications.

Visualizing convergence problems in MCMC simulation. Probabilistic generative modeling is one of the theoretical foundations of current mainstream machine learning and data analysis. Bayesian inference is potentially very powerful but closed-form solutions are seldom available. Inference has to be based on either approximation methods or simulations with Markov Chain Monte Carlo (MCMC) sampling.

The main practical problem of MCMC is how to assess whether the simulation has converged. The resulting samples come from the true distribution only after convergence. It turns out [5] that the main multivariate convergence measure, the multivariate potential scale reduction factor (MPSRF) developed by Brooks and Gelman [6], equals the cost function of a one-dimensional linear discriminant analysis (LDA), a method that discriminates between data classes. Traditional methods of visualizing MCMC simulations do not scale up to large models with lots of parameters. As the cost function of LDA is the equivalent to the MPSRF measure, we can use LDA to focus on features that are relevant to convergence, and thus reduce the number of visualizations.

LDA assumes that each class is normally distributed with the same covariance matrix in each class. This does not hold in general, in particular not before MCMC convergence for small data. To address the above problem, we suggest to complement LDA-based analysis with RCA [5].

Clustering

Discriminative clustering is a method for clustering continuous data so that the clusters become informative of auxiliary data paired with the data samples. While originally motivated by its asymptotic equivalence to vector quantization in learning metrics [7], the current algorithm is a probabilistic algorithm that maximizes the dependency of a contingency table and handles parameter uncertainty in a Bayesian manner [8]. It provides a link between learning metrics and generative modeling.

The key observation is that from the viewpoint of clustering, the parameters defining class distributions within clusters are not interesting and can then be marginalized out. This leads to a cost function only depending on the cluster prototype vectors $\{\mathbf{m}_j\}$:

$$L_{\mathrm{DC}}(\{\mathbf{m}_j\}) \propto \sum_{ij} \log\Gamma(n_i^0 + n_{ji}) - \sum_j \log\Gamma(N^0 + N_j)\ , \tag{6.4}$$

where n_{ji} denotes the number of samples in the cluster j with the value of auxiliary variable $c = i$. The parameters n_i^0 arise from a Dirichlet prior, and $N_j = \sum_i n_{ji}$, $N^0 = \sum_i n_i^0$.

In [8] it was shown that (6.4) can be efficiently optimized by a conjugate gradient algorithm operating on a slightly modified cost function (the counts are smoothed). It gave performance comparable to directly optimizing the original cost function with a time-consuming simulated annealing algorithm. On top of that, the model can be regularized by two alternative ways, either by an information theoretic equalization of cluster sizes, or by a Bayesian way of modeling also the primary data to improve generalization. The latter provides an interesting compromise between modeling conditional and joint densities, and the experimental results show that with small training data sets including a term modeling the covariates improves the accuracy.

Finite-data sequential information bottleneck Count data such as frequencies of words in text documents can be represented as a table of co-occurrence counts, called a contingency table. For clustering such data, successful Information Bottleneck-based methods treat the table as a probability distribution. However, in practice the table is sparsely populated by the finite number of counts, and it is necessary to take the sampling uncertainty into account.

We have introduced a new rigorous method for this. It is a variant of the sequential Information Bottleneck algorithm [9], with a new cost function directly defined for counts. The new cost function, a Bayes factor, compares the posterior probabilities of two alternative probabilistic models for the contingency table. It turns out that we can integrate over model parameters which takes the finite co-occurrence numbers into account. On the other hand, if there is much data, the cost function becomes equivalent to that of the sequential Information Bottleneck.

The new formulation extends discriminative clustering, defined earlier for continuous data, to count data, for which there exist powerful optimization algorithms. The new method, finite sequential Information Bottleneck (fsIB; [10]) outperformed the previous sequential Information Bottleneck in clustering sparse subsets of document corpora, as measured by a precision measure with respect to known categories of the documents.

References

[1] Samuel Kaski and Janne Sinkkonen. Principle of learning metrics for exploratory data analysis. *Journal of VLSI Signal Processing, special issue on Machine Learning for Signal Processing*, 37:177–188, 2004.

[2] Jaakko Peltonen, Arto Klami, and Samuel Kaski. Improved learning of Riemannian metrics for exploratory analysis. *Neural Networks*, 17:1087–1100, 2004.

[3] Jaakko Peltonen and Samuel Kaski. Discriminative components of data. *IEEE Transactions on Neural Networks*, 16(1):68–83, 2005.

[4] Samuel Kaski and Jaakko Peltonen. Informative discriminant analysis. In *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)*, pages 329–336. AAAI Press, Menlo Park, CA, 2003.

[5] Jarkko Venna, Samuel Kaski, and Jaakko Peltonen. Visualizations for assessing convergence and mixing of MCMC. In N. Lavrac, D. Gamberger, H. Blockeel, and L. Todorovski, editors, *Proceedings of the 14th European Conference on Machine Learning (ECML 2003)*, pages 432–443, Berlin, 2003. Springer.

[6] Stephen Brooks and Andrew Gelman. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–456, Dec 1998.

[7] Janne Sinkkonen and Samuel Kaski. Clustering based on conditional distributions in an auxiliary space. *Neural Computation*, 14:217–239, 2002.

[8] S. Kaski, J. Sinkkonen, and A. Klami. Discriminative clustering. *Neurocomputing*, 69:18–41, 2005.

[9] Noam Slonim, Nir Friedman, and Naftali Tishby. Unsupervised document classification using sequential information maximization. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 129–136. ACM Press, New York, NY, USA, 2002.

[10] Jaakko Peltonen, Janne Sinkkonen, and Samuel Kaski. Sequential information bottleneck for finite data. In Russ Greiner and Dale Schuurmans, editors, *Proceedings of the Twenty-First International Conference on Machine Learning (ICML 2004)*, pages 647–654. Omnipress, Madison, WI, 2004.

6.3 Dependency exploration

In dependency exploration we assume that the relevant aspects of data are shared by several information sources. This assumption opens up a new principled framework to combine various information sources. In particular, the effects due to data set-specific noise can be filtered out. Instead of having to specify a model for the noise, the data analyst needs to be able to choose a set of data sources.

The key idea in dependency exploration is to build models for two (or more) information sources such that their statistical dependency is maximized. Given two sets of real-valued vectorial features from two sources, $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^m$, and two representations for them, $v = f_x(\mathbf{x})$ and $w = f_y(\mathbf{y})$, dependency exploration models maximize some estimate of the statistical dependency between v and w. One of the most popular estimators is mutual information $I(V, W) = \sum \sum p(v, w) \log \frac{p(v,w)}{p(v)p(w)}$. The main drawback of mutual information is that it is defined for (known) probability distributions, and may have strong biases when estimated from the finite data of practical applications. We have developed several methods that (i) use mutual information, or (ii) use a finite-data version of mutual information.

Associative clustering

Associative clustering (AC) [1] is a clustering method suitable for dependency exploration of two data sources. It clusters each data source separately, such that the dependency between the clusterings is maximized. If the different sources describe the same object, then the clusterings are as similar as possible in the sense of maximizing statistical dependency, but yet the objects may belong to different clusters in different contexts (sources) if necessary.

The dependencies between the two clusterings are represented with a two-way contingency table formed by cross-tabulation of the data items in the two sets of clusters. AC is an extension of discriminative clustering to two continuous margin spaces, and the techniques similar to discriminative clustering can be applied, including the regularization methods and smoothed partitions. The uncertainty in the clustering result can be dealt with by bootstrap. Figure 6.3 gives on overview of AC.

Generalized canonical correlation analysis

The classical canonical correlation analysis (CCA) finds maximally correlating components from two feature sets. It is equivalent to finding mutual information-maximizing components if the data is normally distributed. A generalized version of CCA, gCCA, extends the method to several feature sets [2].

We have developed a novel way to use the gCCA to integrate multiple data sets in such a way that their statistical dependencies are maximally preserved in the new, integrated representation [3]. As gCCA is computationally efficient it can be used as a preprocessing step for more complicated dependency analysis tasks, such as associative clustering. A sample application is given in Section 5.

Non-parametric dependent components

Canonical correlation analysis can also be extended to capture more general dependencies instead of mere correlation. In [4] we introduced a method coined *dependent component analysis* that estimates the mutual information (or more presicely, the likelihood ratio between dependent and independent hypotheses) using non-parametric density estimates

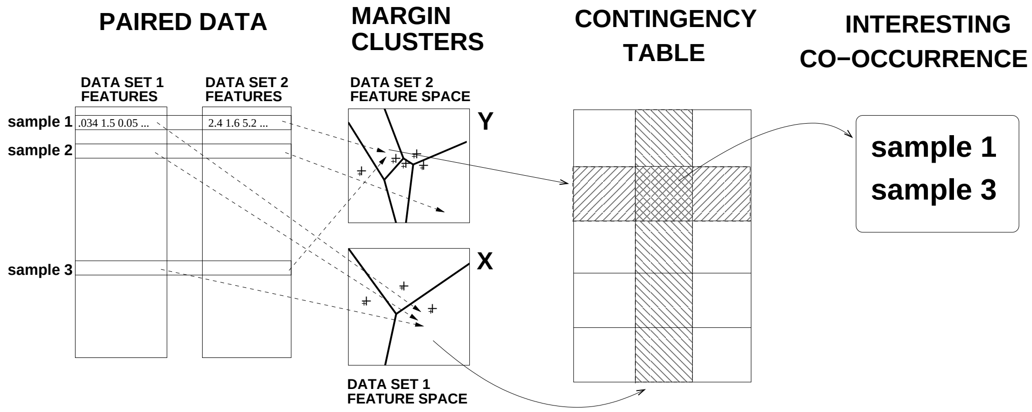


Figure 6.3: Associative clustering (AC) in a nutshell. Two data sets are clustered into margin clusters represented as Voronoi regions with prototype vectors. A one-to-one correspondence between the two data sets exists: each sample is presented in both sets. As each sample falls to one cluster in each data set, we get a contingency table by placing the two sets of clusters as rows and columns, and by counting samples in each combination of row and column clusters. AC by definition finds Voronoi prototypes that maximize the dependency seen in the contingency table. Maximization of dependency in a contingency table results in a maximal amount of counts not explainable by the margin distributions which can then be interpreted as maximally dependent clusters.

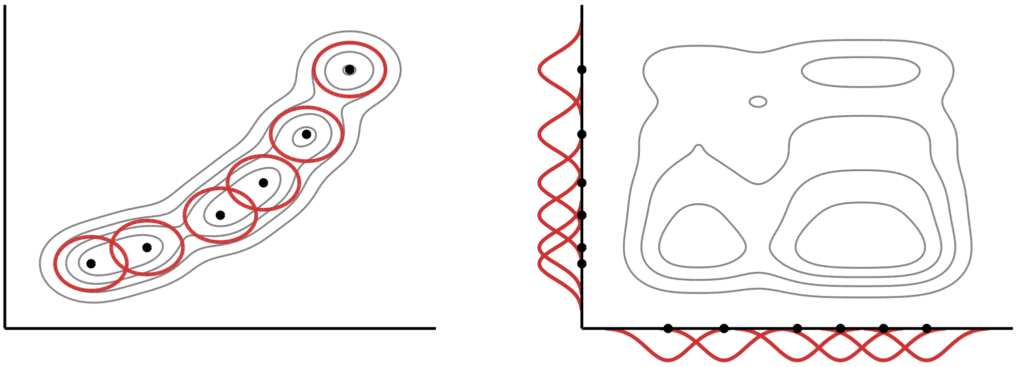


Figure 6.4: Dependency can be measured by studying how closely a joint distribution (left) can be approximated by a product of marginal distributions (right). For independent variables the distributions are identical. Here the gray density estimate in both pictures is computed using Gaussian mixtures, and the representations have relatively clear but non-linear dependency.

in the projection space, and maximizes the estimated dependency with respect to the parameters of the projections using a conjugate gradient algorithm.

The non-parametric measure for dependency is illustrated in Figure 6.4, where the two density estimates required for estimating the likelihood ratio are depicted. The closer the components are to being independent, the closer the two distributions are.

The algorithm was demonstrated to better find the correct component when used on toy data that had dependent non-Gaussian variables. It was also shown that in a real-life application to yeast stress measurements the the projection space found by the method gave more information about general stress than the space found by gCCA or KernelCCA described in [2].

References

[1] Samuel Kaski, Janne Nikkilä, Janne Sinkkonen, Leo Lahti, Juha Knuuttila, and Christophe Roos. Associative clustering for exploring dependencies between functional genomics data sets. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, Special Issue on Machine Learning for Bioinformatics – Part 2*, 2(3):203–216, 2005.

[2] F. R. Bach and M. I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2002.

[3] Janne Nikkilä, Christophe Roos, Eerika Savia, and Samuel Kaski. Explorative modeling of yeast stress response and its regulation with gcca and associative clustering. *International Journal of Neural Systems*, 15(4):237–246, 2005.

[4] Arto Klami and Samuel Kaski. Non-parametric dependent components. In *Proceedings of ICASSP'05, IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages V–209–V–212. IEEE, 2005.

6.4 Discriminative learning

The more traditional counterpart to supervised mining is *discriminative learning* where the data set is the same but the task is different. Given paired data $(\mathbf{x}, c)$, the task is to predict c for a test set where only the values of $\mathbf{x}$ are known.

There exist two traditional modeling approaches for predicting c, discriminative and generative. Discriminative models optimize the conditional probability $p(c|\mathbf{x})$ (or some other discriminative criterion) directly. The models are good classifiers, since they do not waste resources on modeling those properties of the data that do not affect the value of c, that is, the distribution of $\mathbf{x}$. The alternative approach is generative modeling of the joint distribution $p(c, \mathbf{x})$. Generative models add prior knowledge of the distribution of $\mathbf{x}$ into the task. This facilitates for example inferring missing values, since the model is assumed to generate also the covariates $\mathbf{x}$. The models are often additionally simpler to construct, and their parameters offer simple explanations in terms of expected sufficient statistics.

Discriminative Joint Density Models. One way of constructing discriminative classifiers is to take a joint density model, and then change the objective function from the joint likelihood $p(c, \mathbf{x}|\theta)$ to the conditional likelihood $p(c|\mathbf{x}, \theta)$. The obtained solution is (asymptotically) optimal for discrimination [1], given the model family. Compared to pure discriminative models, the benefit of the approach is that prior knowledge about $\mathbf{x}$ can be brought in. We call such a model a discriminative joint density model. The models operate in the same parameter space as ordinary discriminative models, but the generative formulation constrains the model manifold within the space.

Another advantage is that even after converting a joint density model to a discriminative model, the model still constructs a density estimate for $\mathbf{x}$. In [1] we show that this information may be useful, even if the model is inaccurate, for example in predicting missing values of $\mathbf{x}$.

Discriminative Expectation Maximization. Discriminative joint density models have been put to extensive use in speech processing applications, where good results have been obtained using discriminative hidden Markov models [2]. Current state-of-the-art method within the field used for optimizing the models is called extended Baum Welch, EBW [3, 2]. During the last 15 years, considerable effort has been made in order to find the best form of update formulas, but the method is still partly heuristic and has no solid theory explaining the update formulas. In [4] we introduce a discriminative Expectation Maximization -type algorithm that relies on a derivation of a global lower bound for conditional probability densities, alternatively to the practically too complex [5]. The derivation suggests a practical algorithm that results from slightly relaxing the requirement of the globality of the lower bound. The benefit of the algorithm is computational efficiency and simpler update formulas. The resulting update formulas are very close to current extended Baum Welch formulas, for which they give a theoretical basis that justifies the heuristics.

References

[1] Jarkko Salojärvi, Kai Puolamäki, and Samuel Kaski. On discriminative joint density modeling. In João Gama, Rui Camacho, Pavel Brazdil, Alípio Jorge, and Luís Torgo, editors, *Machine Learning: ECML 2005*, Lecture Notes in Artificial Intellligence 3720, pages 341–352, Berlin, Germany, 2005. Springer-Verlag.

[2] D. Povey, P.C. Woodland, and M.J.F. Gales. Discriminative MAP for acoustic model adaptation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP'03)*, volume 1, pages 312–315, 2003.

[3] P.S. Gopalakrishnan, Dimitri Kanevsky, Arthur Nádas, and David Nahamoo. An inequality for rational functions with applications to some statistical estimation problems. *IEEE Transactions on Information Theory*, 37(1):107–113, 1991.

[4] Jarkko Salojärvi, Kai Puolamäki, and Samuel Kaski. Expectation maximization algorithms for conditional likelihoods. In Luc De Raedt and Stefan Wrobel, editors, *Proceedings of the 22nd International Conference on Machine Learning (ICML-2005)*, pages 753–760, New York, USA, 2005. ACM press.

[5] Tony Jebara and Alex Pentland. On reversing Jensen's inequality. In Todd K. Leen, Thomas G. Dietterich, and Volker Tresp, editors, *Advances in Neural Information Processing Systems 13*, Cambridge, MA, April 2001. MIT Press.

6.5 Visualization methods

Visualization of mutual similarities of entries in large data sets is a central subproblem in exploratory analysis and mining. "Looking at the data" may be an invaluable sanity check, and often gives new insights on the data. We have introduced new nonlinear projection methods, and measures for better quantifying the necessary tradeoffs such methods need to make.

Continuity and trustworthiness

We have introduced methods to quantify a key question in visualization, namely the preservation of the original similarity relationships. In general, it is impossible to preserve all the similarities in the data set when projecting it to a lower-dimensional display. Hence, all visualization methods make a compromise between two goals. On the one hand the visualizations should be *trustworthy*, in the sense that samples that are near each other, i.e., in the same neighborhood, in the visualization can be trusted to actually be similar. On the other hand all the original similarities should become visualized. Some of the original similarities might be missing from the visualization because of *discontinuities* in the mapping. The tradeoff between these two goals is quite similar to the precision recall tradeoff in information retrieval.

We compared the trustworthiness and continuity of a set of state-of-the-art methods in a gene expression data visualization task [1, 2]. The Self-organizing map and another nonlinear dimensionality reduction method, Curvilinear Component Analysis, were found to be more trustworthy than other methods while visualizations produced by principal component analysis typically have a good continuity.

Local multidimensional scaling

It would be best if the user could select the tradeoff between trustworthiness and continuity explicitly instead of having to settle for the implicit tradeoff inherent in each method. This idea lead us to develop a new visualization method coined local MDS [3]. It extends a nonlinear dimensionality reduction method, curvilinear component analysis, with the ability to tune the tradeoff.

The cost function of local MDS is

$$\begin{aligned}E = & \frac{1}{2}\sum_i \sum_{j\neq i}[(1-\lambda)(d(\mathbf{x}_i,\mathbf{x}_j)-d(\mathbf{y}_i,\mathbf{y}_j))^2 F(d(\mathbf{y}_i,\mathbf{y}_j),\sigma_i)+ \\ & + \lambda(d(\mathbf{x}_i,\mathbf{x}_j)-d(\mathbf{y}_i,\mathbf{y}_j))^2 F(d(\mathbf{x}_i,\mathbf{x}_j),\sigma_i)]\ .\end{aligned}$$

The first part of the cost function focuses on preserving distances that are within the area of influence $F(d(\mathbf{y}_i,\mathbf{y}_j),\sigma_i)$ around a data point i in the output space, and is the same as the original cost function of curvilinear component analysis. The second part focuses on preserving distances that are within the area of influence in the input space. The weighting between these two parts, and the tradeoff between trustworthiness and continuity, is controlled with the parameter λ. The effect of changing λ is illustrated in Figure 6.5.

References

[1] Samuel Kaski, Janne Nikkilä, Merja Oja, Jarkko Venna, Petri Törönen, and Eero Castrén. Trustworthiness and metrics in visualizing similarity of gene expression. *BMC Bioinformatics*, 4:48, 2003.

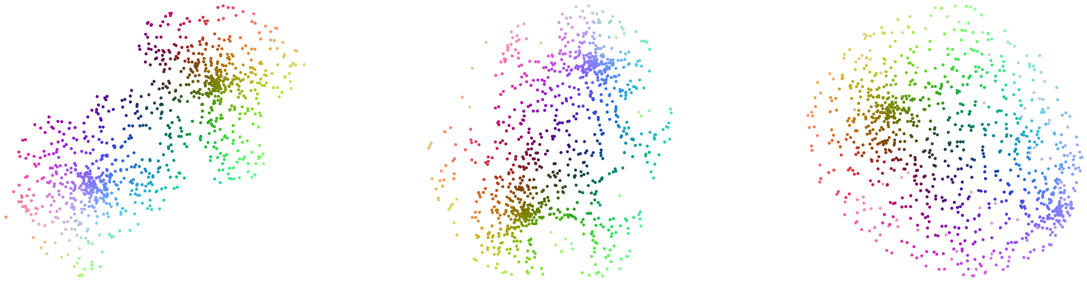


Figure 6.5: Three projections of a three-dimensional spherical cell with local MDS. On the left, trustworthiness of the projection is maximized by selecting $\lambda = 0$. In the middle and right, discontinuity of the projection is penalized as well, by setting $\lambda = 0.1$ and $\lambda = 0.9$, respectively. When λ is increased the edges where continuity is violated the worst get pulled closer together to minimize the number of neighborhoods that become split, and to reduce the distance between those neighborhoods that cannot be connected.

[2] Jarkko Venna and Samuel Kaski. Visualized atlas of a gene expression databank. In *Proceedings of Symposium of Knowledge Representation in Bioinformatics*, pages 30–36, Espoo, Finland, 2005.

[3] Jarkko Venna and Samuel Kaski. Local multidimensional scaling with controlled tradeoff between trustworthiness and continuity. In *Proceedings of 5th Workshop on Self-Organizing Maps*, pages 695–702, Paris, France, 2005.

Chapter 7

Image retrieval and analysis

Erkki Oja, Jorma Laaksonen, Jukka Iivarinen, Markus Koskela, Ramūnas Girdziušas, Jussi Pakkanen, Ville Viitaniemi, Zhirong Yang, Rami Rautkorpi, Mats Sjöberg, Hannes Muurinen

7.1 Content-based image retrieval by self-organizing maps

Erkki Oja, Jorma Laaksonen, Markus Koskela, Ville Viitaniemi, Zhirong Yang, Mats Sjöberg, Hannes Muurinen

Content-based image retrieval (CBIR) has been a subject of intensive research effort for more than a decade now. It differs from many of its neighboring research disciplines in computer vision due to one notable fact: human subjectivity cannot totally be isolated from the use and evaluation of CBIR systems. In addition, two more points make CBIR systems special. Opposed to such computer vision applications as production quality control systems, operational CBIR systems would be very intimately connected to the people using them. Also, effective CBIR systems call for means of interchanging information concerning images' content between local and remote databases, a characteristic very seldom present, e.g., in industrial computer vision.

PicSOM

The methodological novelty of our neural-network-based CBIR system, PicSOM [1, 2], is to use several Self-Organizing Maps in parallel for retrieving relevant images from a database. These parallel SOMs have been trained with separate data sets obtained from the image data with different feature extraction techniques. The different SOMs and their underlying feature extraction schemes impose different similarity functions on the images. In the PicSOM approach, the system is able to discover those of the parallel SOMs that provide the most valuable information for each individual query instance.

Instead of the standard SOM version, PicSOM uses a special form of the algorithm, the Tree Structured Self-Organizing Map (TS-SOM) [3]. The hierarchical TS-SOM structure is useful for large SOMs in the training phase. In the standard SOM, each model vector has to be compared with the input vector in finding the best-matching unit (BMU). With the TS-SOM one follows the hierarchical structure which reduces the complexity of the search to $O(\log n)$. After training each TS-SOM hierarchical level, that level is fixed and each neural unit on it is given a visual label from the database image nearest to it.

Self-organizing relevance feedback

When we assume that similar images are located near each other on the SOM surfaces, we are motivated to exchange the user-provided relevance information between the SOM units. This is implemented in PicSOM by low-pass filtering the map surfaces. All relevant images are first given equal positive weight inversely proportional to the number of relevant images. Likewise, nonrelevant images receive negative weights that are inversely proportional to their number. The relevance values are then summed in the BMUs of the images and the resulting sparse value fields are low-pass filtered.

Figure 7.1 illustrates how the positive and negative relevance responses, displayed with red and blue map units, respectively, are first mapped on a SOM surface and how the responses are expanded in the low-pass filtering. As shown on the right side of the figure, the relative distances of SOM model vectors can also be taken into account when performing the filtering operation [4]. If the relative distance of two SOM units is small, they can be regarded as belonging to the same cluster and, therefore, the relevance response should easily spread between the neighboring map units. Cluster borders, on the other hand, are characterized by large distances and the spreading of responses should be less intensive.

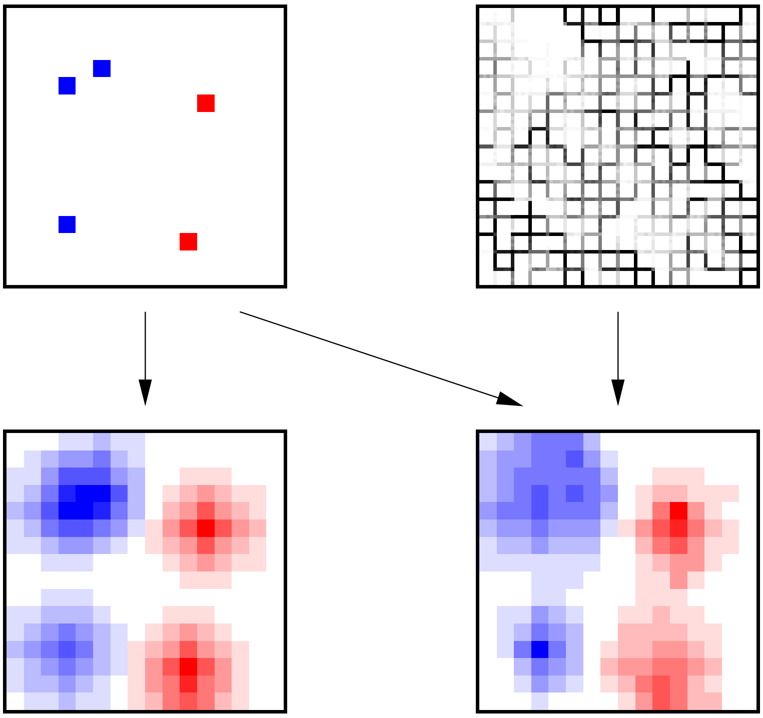


Figure 7.1: An example of how positive and negative map units, shown with red and blue marks on the top-left figure, are low-pass filtered. Two alternative methods exist; either we ignore (bottom-left figure) or take into account (bottom-right figure) the relative distances between neighboring SOM model vectors. In the top-right figure, the relative distances are illustrated with gray level bars so that a darker shade of gray corresponds to a longer relative distance between two neighboring map units.

Finally, the set of images forming the result of the query round is obtained by summing the relevance responses or *qualification values* from all the used SOMs. As a result, the different content descriptors do not need to be explicitly weighted as the system automatically weights their opinions regarding the images' similarity and relevance.

User interaction feature

Relevance feedback can be seen as a form of supervised learning to adjust subsequent query rounds by using information gathered from the user's feedback. It is essential that the learning takes place during one query, and the results are erased when starting a new one. This is because the target of the search usually changes from one query to the next, and so the previous relevances have no significance any more. This is therefore *intra-query* learning.

Relevance feedback provides information which can also be used in an *inter-query* or *long-term* learning scheme. The relevance evaluations provided by the user during a query session partition the set of seen images into relevant and nonrelevant classes with respect to that particular query target. The fact that two images belong to the same class is a cue for similarities in their semantic content. This information can be utilized by considering the previous user interaction as metadata associated with the images and use it to construct a *user interaction* or *relevance feature*, to be used alongside with the visual features. This method was presented and experimented with in [5]. An example of a resulting SOM is illustrated in Figure 7.2. In the figure, a 16×16-sized SOM trained with user interaction

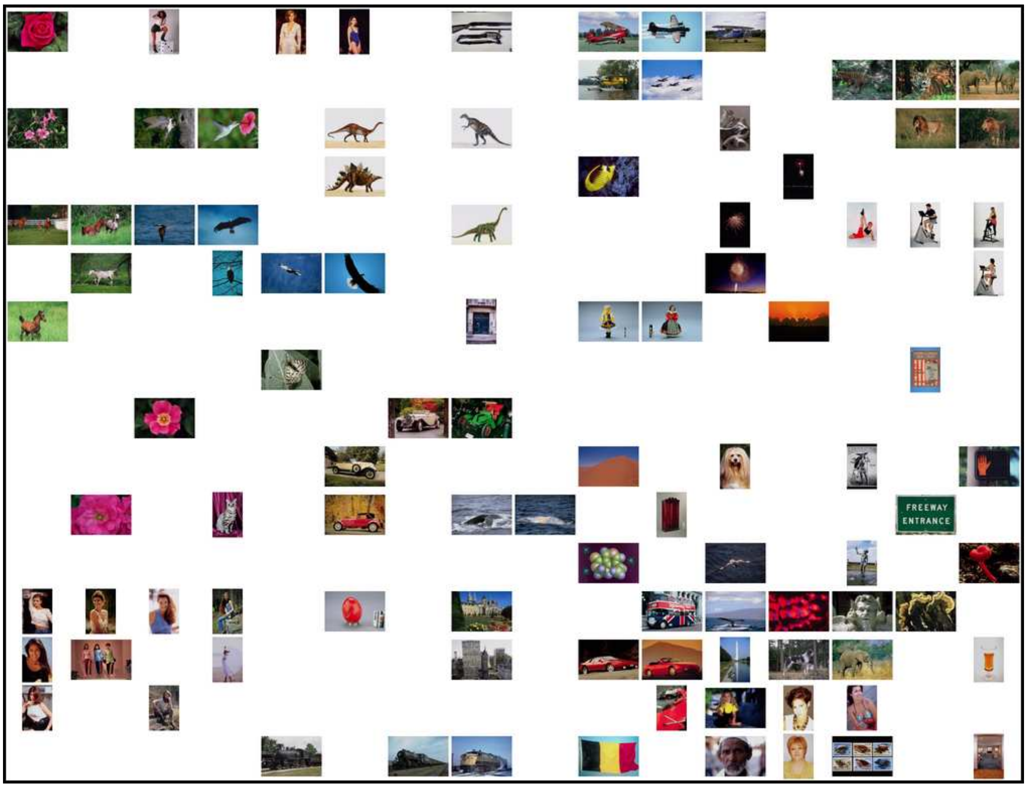


Figure 7.2: The image labels of a 16×16-sized SOM trained with user interaction data.

data is shown. It can be observed that images with similar semantic content have been mapped near each other on the map.

In some cases, the image database may also contain manually assigned or implicit annotations. These annotations describe high-level semantic content of the image and often contain invaluable information for retrieval purposes. Therefore, it is useful to note that the user-provided relevance evaluations discussed above are notably similar to these annotations. In particular, keyword annotations can be seen as high-quality user assessments and the presented method can be readily utilized also for these annotations.

Use of segmented images

The general problem of image understanding is intrinsically linked to the problem of image segmentation. That is, if one understands an image one can also tell what the different parts of it are. Segmentation thus seems to be a natural part of image understanding. As the objects and entities in the images usually hierarchically decompose into sub-objects, the segmentations defined by the images form hierarchical segmentation trees.

For an automatic system segmenting an image is never trivial and the results seldom correspond to the real objects in the picture. But even so segmentation may be useful in CBIR, because different, visually homogeneous regions somehow characterize the objects and scenes in the image. The existing approaches differ mainly in the fashion the segment-wise similarities are combined to form image-wise similarities used in the retrieval. The hierarchical structure of segmentations is usually ignored.

The PicSOM system uses the results of unsupervised segmentation by generalizing the original algorithm so that not only the entire images but also the image segments are seen as objects in their own right. The segments are also considered to be sub-objects of the images they are a part of. The relevance feedback process is modified so that when an image is marked as relevant all its sub-objects (segments) are also marked as relevant. Then, after calculating qualification values for all the objects on the different TS-SOMs, the qualification values of all the sub-objects are summed to their parent objects. Finally, the values obtained from different maps are again summed up to form the final image-wise qualification values. In the algorithm, different levels of segmentation hierarchies can be considered simultaneously. The approach can be used also for salient image parts that have been identified by detecting interest points in the images.

In practical applications and preliminary experiments we have observed that for most of the used ground truth image classes, the retrieval precision obtained by using both entire and segmented images together excels over that obtained by using either ones alone. In forthcoming systematic experiments we will further explore the usefulness of these various approaches for exploiting the spatial decomposition of the images into subparts on one hand, and combining the approaches on the other hand.

In addition to interactive CBIR, the PicSOM system has been used off-line to investigate the connection between semantic concepts and results of unsupervised segmentation. In particular, we have considered the keyword focusing problem where the system autonomously learns to attribute annotating keywords to specific segments. The system discovers the segment-keyword connections from example images where the annoting keywords are given on per-image basis. An example of a relevance focusing setting is shown in Figure 7.3 where pieces of clothing are localised in a database of fashion images.

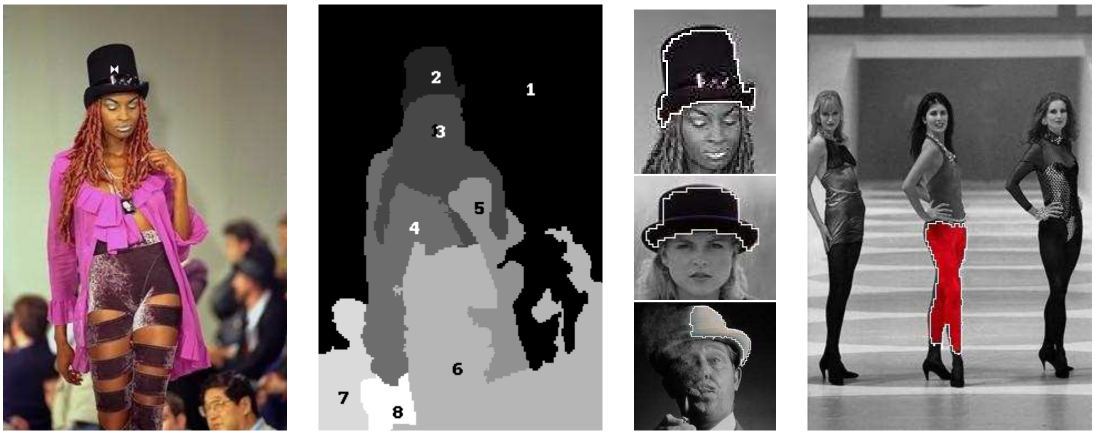


Figure 7.3: Focusing keywords “hat” and “red trousers”.

Interactive facial image retrieval

Most existing face recognition systems require the user to provide a starting image, which is however not practical e.g. when searching for a criminal based on a witness' power of recall. Interactive facial image retrieval, which are mainly based on learning the relevance feedback from the user, is a potential approach to address this problem.

The early appearance of the first subject hit is critical for the success of the retrieval. Unlike content-based image retrieval systems based on general images, the query precision on facial images suffers from the problem of extremely small sizes of the *subject classes* [6]. If only images that depict the correct person are regarded as relevant, many pages of only non-relevant images would be displayed. Because the negative responses from the user in the early rounds provide little semantic information, the iteration progresses in a

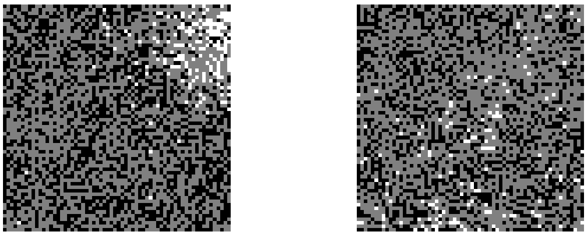


Figure 7.4: A 64 × 64 DSOM example shown on the left. The white points represent the class members of *mustache_yes* and black points for *mustache_no*. The normal SOM without discriminative preprocessing is shown on the right for comparison.

nearly random manner.

We have proposed a novel method which adaptively learns partial relevance during the interactive retrieval. First, we extended the PicSOM CBIR system by replacing the membership of a subject class with that of a semantic class as the new relevance criterion until the first subject hit appears. Second, we applied supervised learning as a preprocessing step before training the Self-Organizing Maps so that the resulting SOMs have stronger discriminative power. Figure 7.4 visualizes an example *Discriminative Self-Organizing Map* (DSOM) of two semantic classes, *mustache_yes* and *mustache_no*. The empirical results on FERET database [7] have shown that the number of displayed images can be significantly reduced by employing these two strategies [8].

Different ways for obtaining a DSOM exist. We applied Fisher's linear discriminant analysis (LDA) as a preprocessing step before training the SOMs. More advanced methods for extracting the discriminative components of data are now under development.

Multimodal hierarchical objects

The basic ideas of CBIR can be expanded to the more general concept of content-based *information* retrieval if we also consider objects of other types than images, for example text, video, and multi-media objects in general. The only restriction is that we must be able to extract some kind of low-level statistical feature vectors from the new data objects. For example, text data similarity can be evaluated with the n-gram method.

A possibly even more important development has been the incorporation of the context and relationships of data items as hierarchical objects in PicSOM. For example a web page with text, embedded images and links to other web pages can be modeled as a hierarchical object tree with a web page object as parent and the text, links and images as children.

The PicSOM system has been extended to support such general multimodal hierarchical objects and a method for *relevance sharing* between these objects has been implemented [9]. This means that the relevance assessments originally received from user feedback will be transferred from the object to its parents, children and siblings. This means for example, if we want to search for an image of a cat from a multimedia message database, we can let the system compare not only the images but also the related textual objects. If the reference message text contains the word "cat" we can find images which are not necessarily visually similar, but have related texts containing the same keyword.

The hierarchical object paradigm has been applied to many problems, such as web-page structures and e-mail message retrieval. Most recently the PicSOM group took part in the NIST TRECVID 2005 workshop where we successfully applied these methods to video retrieval [10]. In addition to the hierarchical object method we also applied *class models* as a way of representing semantic concepts. The class models were previously used for image group annotation in [11], whereas in the TRECVID context we are interested in video shots of the test collection that have the highest likelihood of being relevant to a given concept.

The multi-part hierarchy for video shots used for indexing the TRECVID 2005 collection is illustrated in Fig. 7.5. The video shot itself is considered as the main or parent object. The keyframes (one or more) associated with the shot, the audio track, and ASR/MT text are linked as children of the parent object. All object modalities may have one or more SOMs or other feature indices, and thus all objects in the hierarchy may have links to a set of associated feature indices.

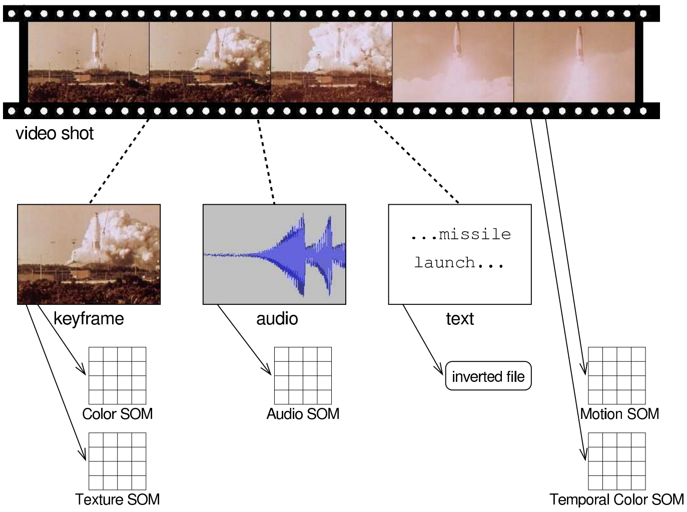


Figure 7.5: The hierarchy of video and multimodal SOMs.

Class models were created by utilizing semantic classifications of the training set videos provided by NIST, such as a list of clips showing an explosion or fire. Feature vectors of objects belonging to such class models can be mapped as impulses onto to the corresponding SOM surfaces, providing an estimate of the true class distribution. In retrieval the sign of the impulses were adjusted to represent relevant (positive) and non-relevant (negative) concepts. The sparse value fields on the maps are low-pass filtered to spread the information which increases the probability that objects mapped to nearby SOM units will be retrieved.

In the high-level feature extraction task we applied the class model representations to find video clips representing different concepts, applying an SFS-type feature selection scheme for each concept separately. The search tasks were run with the PicSOM system by using predefined search topics in three ways: with user interaction, with some manual tweaking of the parameters and without user interaction. To our delight, the PicSOM results compared very well with the other systems taking part in the TRECVID 2005 evaluation. For example, in the automatic search tasks, the PicSOM results were clearly above the median in all but one topic.

References

[1] J. T. Laaksonen, J. M. Koskela, S. P. Laakso, and E. Oja. PicSOM - Content-based image retrieval with self-organizing maps. *Pattern Recognition Letters*, 21(13-14):1199–1207, November 2000.

[2] J. Laaksonen, M. Koskela, and E. Oja. PicSOM – Self-Organizing Image Retrieval with MPEG-7 Content Descriptors. *IEEE Transactions on Neural Networks*, 13(4): 841-853, July 2002.

[3] P. Koikkalainen and E. Oja. Self-organizing hierarchical feature maps. In *Proc. International Joint Conference on Neural Networks*, vol. II, pages 279-285, Piscataway, NJ, 1990.

[4] M. Koskela, J. Laaksonen, and E. Oja. Implementing Relevance Feedback as Convolutions of Local Neighborhoods on Self-Organizing Maps. In *Proc. International Conference on Artificial Neural Networks*, pages 981-986. Madrid, Spain. August 2002.

[5] M. Koskela and J. Laaksonen. Using Long-Term Learning to Improve Efficiency of Content-Based Image Retrieval. In *Proc. Third International Workshop on Pattern Recognition in Information Systems*, pages 72-79, Angers, France, April 2003.

[6] Zhirong Yang and Jorma Laaksonen. Interactive retrieval in facial image database using Self-Organizing Maps. In *Proceedings of IAPR Conference on Machine Vision Applications (MVA 2005)*, pages 112–115, Tsukuba Science City, Japan, May 2005.

[7] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss. The FERET evaluation methodology for face recognition algorithms. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22:1090–1104, October 2000.

[8] Zhirong Yang and Jorma Laaksonen. Partial relevance in interactive facial image retrieval. In *Proceedings of 3rd International Conference on Advances in Pattern Recognition (ICAPR 2005)*, pages 216–225, Bath, UK, August 2005.

[9] Mats Sjöberg and Jorma Laaksonen. Content-based retrieval of web pages and other hierarchical objects with Self-Organizing Maps. In *Proceedings of 15th International Conference on Artificial Neural Networks (ICANN 2005)*, pages 841–846, Warsaw, Poland, September 2005.

[10] Markus Koskela, Jorma Laaksonen, Mats Sjöberg, and Hannes Muurinen. PicSOM experiments in TRECVID 2005. In *Proceedings of the TRECVID 2005 Workshop*, pages 262–270, Gaithersburg, MD, USA, November 2005.

[11] Markus Koskela and Jorma Laaksonen. Semantic annotation of image groups with Self-Organizing Maps. In *Proceedings of 4th International Conference on Image and Video Retrieval (CIVR 2005)*, pages 518–527, Singapore, July 2005.

7.2 Content-based retrieval of defect images

Jussi Pakkanen, Jukka Iivarinen, Rami Rautkorpi

A need for efficient and fast methods for content-based image retrieval (CBIR) has increased rapidly during the last decade. The amount of image data that has to be stored, managed, browsed, searched, and retrieved grows continuously on many fields of industry and research.

In this project we have taken a noncommercial CBIR system called PicSOM, and applied it to several databases of surface defect images. PicSOM has been developed in our laboratory at Helsinki University of Technology to be a generic CBIR system for large, unannotated databases. We have made some modifications to the original PicSOM system that affect mostly feature extraction and visualization parts of PicSOM. As an extra problem-specific knowledge we have segmentation masks for each defect image. This information is utilized in PicSOM so that feature extraction is only done for defect areas in each defect image.

Overview of the method

Interpretation of defect images is a demanding task even to an expert. The defect images concerned in this work contain surface defects, and they were taken from a real, online process. Currently we have two major database types: paper and metal surface defects. Both of these types contain several different defect classes (e.g. dark and light spots, holes, scratches, oli stains and so on) that are fuzzy and overlapping, so it is not possible to label defects unambiguously.

In the present work we have adopted the PicSOM system as our content-based image retrieval (CBIR) system and embedded the defect image databases into PicSOM. PicSOM has several features that make it a good choice for our purposes. The most important of these is the fact, that PicSOM can effectively combine search results of different features. This makes adding new features fast and efficient.

Features for defect characterization Several types of features can be used in PicSOM for image querying. These include features for color, shape, texture, and structure description of the image content. When considering defect images, there are two types of features that are of interest: shape features and internal structure features. Shape features are used to capture the essential shape information of defects in order to distinguish between differently shaped defects, e.g. spots and wrinkles. Internal structure features are used to characterize the gray level and textural structure of defects.

One of the advantages of PicSOM is its open architecture. This makes it simple to add new features to the system. Originally we used simple descriptors for shape, texture features based on the co-occurrence matrix, and the gray level histogram. We use the following set, which consists mainly of features from the MPEG-7 standard, with some additional features that we have developed for the shape description of surface defects.

Scalable Color descriptor is a 256-bin color histogram in HSV color space, which is encoded by a Haar transform.

Color Layout descriptor specifies a spatial distribution of colors. The image is divided into 8 × 8 blocks and the dominant colors are solved for each block in the YCbCr color system. Discrete Cosine Transform is applied to the dominant colors in each channel and the DCT coefficients are used as a descriptor.

Color Structure descriptor captures both color content and the structure of this content. It does this by means of a structuring element that is slid over the image. The numbers of positions where the element contains each particular color is recorded and used as a descriptor. As a result, the descriptor can differentiate between images that contain the same amount of a given color but the color is structured differently.

Edge Histogram descriptor represents the spatial distribution of five types of edges in 16 sub-images. The edge types are vertical, horizontal, 45 degree, 135 degree and non-directional, and they are calculated by using 2×2-sized edge detectors for the luminance of the pixels. A local edge histogram with five bins is generated for each sub-image, resulting in a total of 80 histogram bins.

Homogeneous Texture descriptor filters the image with a bank of orientation and scale tuned filters that are modeled using Gabor functions. The first and second moments of the energy in the frequency domain in the corresponding sub-bands are then used as the components of the texture descriptor.

Shape feature For shape description, we use our own problem-specific shape feature set that was developed for surface defect description. It consists of several simple descriptors calculated from a defect's contour. The descriptors are convexity, principal axis ratio, compactness, circular variance, elliptic variance, and angle.

Edge Co-occurrence Matrix was developed for enhanced shape description. It is similar to the Gray Level Co-occurrence Matrix, but instead of a gray level intensity image, it is computed from an edge image. In this respect it is related to the Edge Histogram, but the entire image is processed at once and eight edge directions at 45 degree intervals are detected. The result is an 8×8 matrix representing the frequencies of edge direction pairs that have a given displacement.

Contour Co-occurrence Matrix was developed in conjunction with the Edge Co-occurrence Matrix. It is computed from the chain-code representation of the contour of a defect. Each link in the chain is a step between two successive pixels on the contour, with eight possible link directions. The result is an 8×8 matrix representing the frequencies of link direction pairs that have a given displacement.

These features were found to work very well on classification experiments using a smaller, pre-classified data base. Similar or better performance is obtained when the Edge Co-occurrence Matrix and the Contour Co-occurrence Matrix are substituted for the Edge Histogram descriptor and the earlier shape feature set.

Experiments

The problem at hand is now the following one: Given a new defect or a set of defects, retrieve similar defects that might have appeared previously. The retrieval is based on shape and internal structure features, so there is no need for manual annotation or labeling. The largest defect database has almost 45000 defect images that were taken from a real, online process. The images have different kinds of defects, e.g. dark and light spots, holes, and wrinkles. They are automatically segmented beforehand so that each defect image has a gray level image and a binary segmentation mask that indicates defect areas in the image. The image database was provided by our industrial partner, ABB oy.

Two example queries in Figure 7.6 show that the system works quite well. Under the TS-SOMs are the images selected by the user (the so called query images), and at the

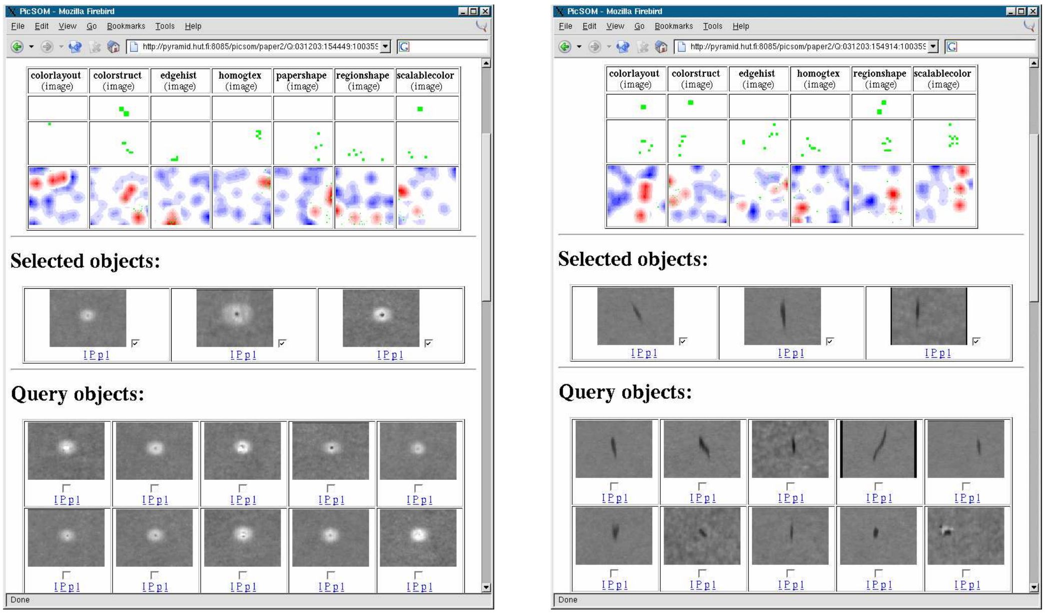


Figure 7.6: Example PicSOM queries.

bottom are the images returned by the PicSOM system. All returned images are visually similar to the query images. The system retains a similar level of success when queried with different types of defects. The true power comes from combining the maps. The PicSOM engine combines the various maps in a powerful manner, yielding good results.

Conclusions

In this project a noncommercial content-based image retrieval (CBIR) system called PicSOM is applied to retrieval of defect images. New feature extraction algorithms for shape and internal structure descriptions are implemented in the PicSOM system. The results of experiments with almost 45000 surface defect images show that the system works fast with good retrieval results.

References

[1] J. Iivarinen and J. Pakkanen, Content-Based Retrieval of Defect Images, In *Proceedings of Advanced Concepts for Intelligent Vision Systems*, pp. 62–67, 2002.

[2] J. Pakkanen and J. Iivarinen, Content-based retrieval of surface defect images with MPEG-7 descriptors, In K. Tobin Jr. and F. Meriaudeau, editors, *Proceedings of Sixth International Conference on Quality Control by Artificial Vision*, Proc. SPIE 5132, pp. 201–208, 2003.

[3] R. Rautkorpi, J. Iivarinen, Shape-Based Co-occurrence Matrices for Defect Classification, In *Proceedings of 14th Scandinavian Conference on Image Analysis*, pp. 588–597, 2005.

7.3 Alterable volume flow in the use of input deformations for a massive Gaussian process smoothing

Ramūnas Girdziušas, Jorma Laaksonen

The use of input deformations with nonlinear regression models is relatively unpopular for a reason that they introduce another degree of nonlinearity into the model identification. However, in a massive regression with low-dimensional inputs (d=1,2,3) they have a visual representation and enhance the predictive power of simple stationary isotropic homogeneous models. Considering Gaussian process regression, input deformations are linked directly to the smoothing properties of the covariance functions:

$$\mathrm{Cov}\,[f(\mathbf{x}_i), f(\mathbf{x}_j)] \;\;\mapsto\;\; \mathrm{Cov}[f(\mathbf{x}_i - \mathbf{u}(\mathbf{x}_i)), f(\mathbf{x}_j - \mathbf{u}(\mathbf{x}_j))]. \tag{7.1}$$

A direct maximum likelihood estimation of the displacement field $\mathbf{u}(\mathbf{x})$ is hardly possible because it is unclear what smoothness priors to utilize and how to maintain the model space within the computationally feasible level.

We are investigating a likelihood-ascending nonparametric estimation of displacements which allow to alter the deformation volume and is computationally feasible. A suggestion is to utilize covariance functions of a univariate Brownian motion and extend them to a higher-dimensional case via an additive operator splitting [4]. Input deformations of such a model do not alter the structure of the inverse covariance matrix. It will remain tridiagonal [1] which makes the evaluation of the log-likelihood particularly easy. The Gaussian process log-likelihood $\mathcal{L}_u$ can then be maximized by embedding the displacement field into a simplified fluid flow [3]. This amounts to performing a non-gradient descent on $-\mathcal{L}_u + \mathbf{R}_{\lambda,\mu}$, where the Lamé functional $\mathbf{R}_{\lambda,\mu}$ is a positive definite quadratic form of the spatial derivatives of the displacements. However, it does not act as a regularization functional:

$$\mathcal{L}(\mathbf{u}_t + \mathbf{du}_t) \approx \mathcal{L}(\mathbf{u}_t) + \nabla_u \mathcal{L}(\mathbf{u}_t)^T \mathbf{du}_t, \tag{7.2}$$

$$\mathcal{R}_{\lambda,\mu}(\mathbf{u}_t + \mathbf{du}_t) \mapsto \mathcal{R}_{\lambda,\mu}(\mathbf{u}_t) + \mathbf{du}_t^T \nabla_u \nabla_u^T \mathcal{R}_{\lambda,\mu}(\mathbf{u}_t) \mathbf{du}_t \;, \tag{7.3}$$

The constants λ and μ define the compressibility of the flow. This approach has been tested in a synthetic problem of a bivariate regression where it is equally competitive with a robust but not visual regression techniques [3] and in the estimation of an optical flow, which does not reach the state of the art but is not dramatically far from it [2].

References

[1] R. Girdziušas and J. Laaksonen. Use of input deformations with Brownian motion filters for discontinuous regression. In *ICAPR*, volume 3686 of *LNCS*, p. 219–228. Springer, 2005.

[2] R. Girdziušas and J. Laaksonen. Optimal ratio of Lamé moduli with application to motion of Jupiter storms. In *SCIA*, vol. 3540 of *LNCS*, p. 1096–1106. Springer, 2005.

[3] R. Girdziušas and J. Laaksonen. Gaussian process regression with fluid hyperpriors. In *ICONIP*, vol. 3316 of *LNCS*, pages 567–572. Springer, 2004.

[4] B. Fischer and J. Modersitzki. *Inverse Problems, Image Analysis, and Medical Imaging*, volume 313 of *AMS Contemporary Mathematics*, chapter Fast Diffusion Registration, pages 117–129. 2002.

7.4 Stopping criteria for nonlinear diffusion filters

Ramūnas Girdziušas, Jorma Laaksonen

Nonlinear diffusion filters preserve discontinuities without requiring a prior knowledge of their approximate locations and reveal the optimal basis of a signal in an unsupervised manner [1]. Potential applications include a fast retrospective multiple change-point detection in nuclear magnetic response measurements of well log data, restoration of aerial images of buildings, fingerprint enhancement and optical flow.

We consider a nonlinear variation-diminishing diffusion of an image $y(\mathbf{x})$, defined by

$$\partial_t f = \nabla \cdot \left(\phi_\theta(||\nabla f||)\nabla f\right), \ \ s.t. \ \nabla f|_\Omega = 0, \tag{7.4}$$

with $f \equiv f_t(\mathbf{x})$, $f_0(\mathbf{x}) = y(\mathbf{x})$, $\mathbf{x} \in \Omega = [0,1]^d$, $d = 1, 2$. The univariate diffusivity function $\phi_\theta(z) \geq 0$ depends on a few parameters $\boldsymbol{\theta}$ and it is such that $\phi'_\theta(z)/(z)$ is continuous and decreasing.

Eq. (7.4) produces a scale-space of images and needs systematic criteria for the selection of the optimal stopping time. A possible remedy to this problem is to extend Eq. (7.4) in order to define a joint probability measure for the observations, diffusion outcome and the optimal stopping time, which could be accomplished within the framework of Itô diffusions. This turns out to be a formidable task.

Among a variety of heuristic stopping principles proposed in the last decades, one of the simplest and the most stable criterion stops the diffusion when the filtering outcome becomes uncorrelated with the noise estimate between the observations and a true signal [4]. Motivation for the decorrelation is based on the theory of a linear diffusion.

We have shown [1] that the decorrelation principle is a particular case of a maximum likelihood estimation in an additive white Gaussian noise when the assumption of a true signal is the Gaussian process defined by

$$F_t \sim \mathcal{GP}\left(0, \theta_0 \left(T_\theta - Id\right)^{-1}\right), \tag{7.5}$$

where θ_0 is a known constant, T_θ denotes the operator that maps an initial image to the diffusion outcome, it depends on the unknown diffusivity parameters $\boldsymbol{\theta}$ and the stopping time t. Such a model is useful because: (i) it explains why decorrelating criterion can be negative in nonlinear diffusion and gives its precise bounds, (ii) shows that the decorrelation principle overestimates the optimal stopping time, and (iii) provides better stopping criteria [2, 3].

References

[1] R. Girdziušas and J. Laaksonen. Gaussian processes of nonlinear diffusion filtering. In *IJCNN*, vol. 2, p. 1012–1017. IEEE, 2005.

[2] R. Girdziušas and J. Laaksonen. Jacobi alternative to Bayesian evidence maximization in diffusion filtering. In *ICANN*, vol. 3697 of *LNCS*, p. 247–252. Springer, 2005.

[3] R. Girdziušas and J. Laaksonen. Optimal stopping and constraints for diffusion models of signals with discontinuities. In *ECML*, vol. 3720 of *LNAI*, p. 576–583. Springer, 2005.

[4] P. Mrázek and M. Navara. Selection of optimal stopping time for nonlinear diffusion filtering. *Int. Journal of Computer Vision*, 52(2):189–203, 2003.

Chapter 8

Adaptive cognitive systems

Timo Honkela, Krista Lagus, Ville Könönen, Ann Russell, Mikaela Klami, Tiina Lindh-Knuutila, Matti Pöllä, Jaakko Väyrynen, Kevin Hynnä

8.1 Introduction

Our research on cognitive systems focuses on modeling and applying methods of unsupervised and reinforcement learning. The general aim is to provide a methodological framework for theories of conceptual development, symbol grounding, communication among autonomous agents, agent modeling, and constructive learning. We also work in close collaboration with other groups in our laboratory, e.g., related to multimodal environments.

An important part of our acitivity has been the active role in organizing international scientific events:

- International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05) [2],
- Symposium on Adaptive Models of Knowledge, Language and Cognition (AMKLC'05) [12], organized in conjunction with AKRR'05 conference, and
- Workshop on Reinforcement Learning in Non-Stationary Environments in conjunction with the 16th European Conference on Machine Learning (ECML'2005) [9].

International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning, AKRR'05, was raising awareness of adaptive approaches to knowledge representation and reasoning. The power of the adaptive systems lies in the fact that they enable computers to adapt to the needs of individuals, groups, enterprises and organizations in the changing world. There were two special symposia in the conference that provide a focused view on their topics: Adaptive Models of Knowledge, Language and Cognition (AMKLC'05) and Knowledge Representation for Bioinformatics (KRBIO'05). The goal of the reinforcement learning workshop in ECML'2005 was to foster co-operation inside European reinforcement learning research community and raise international visibility of European reinforcement learning research.

8.2 Emergence of cognitive and conceptual representations

Conceptual modeling is a task which has traditionally been conducted manually. In artificial intelligence, knowledge engineers have written descriptions of various domains using formalisms based on predicate logic and other symbolic representations such as semantic networks and rule-based systems. As modern related topics, the Semantic Web and knowledge representation formalisms like eXtendable Markup Language (XML) can be mentioned.

Philosophical considerations

The traditional symbolic approach has concentrated on the linguistic domain. Therefore, the models often lack the connection to the perceptual domain. It has been assumed that knowledge can be represented as propositional structures that are based on static shared concepts. It has been commonplace to assume that there is a rather one-to-one correspondence between words and concepts. Moreover, it is assumed that a concept refers unambiguously to a number of distinct objects or events in the reality. The individual differences are assumed to be small and explained as errors.

In radical constructivism (consider, e.g., [8, 15]), it is pointed out that cognitive agents construct their description of the world, and this description consists of constructed categories such as objects and events along with their associated subcategories. Each of those constructions is subjective but at the same time their formation is based on the interaction with other agents as well as artefacts that reflect the structural characteristics of the constructions of other agents. It should not be be taken as a fact that only the rules or principles observed in the past shall apply to the future. Constructive learning involves qualitative restructuring and modification of internal knowledge representations, rather than just accumulation of new information in memory. [5]

The static nature of the information systems in general makes them also prone to be “incompatible with the reality”. One reason is that the domain of use is changing. Another, more profound reason is that human beings have individual conceptual systems, gained through constructive learning processes. A conceptually static and coarse-grained information system matches with our conceptual systems only partially. This misfit may lead to errors or unjustified procedures. Therefore, it appears necessary that any information system should be adaptive in order to be able to deal with the variety of conceptual construction and in order to be able to conduct meaning negotiations.[3]

Emergence of a shared conceptual system

We studied the emergence of associations between concepts and words. The important questions being how a language learner, or an agent, learns the meaning of new words, and how an agreement on the use of words is reached in a community of agents. The Self-Organizing Map (SOM) were used as a model of an agent's conceptual map, and concepts are seen as areas formed in a SOM based on unsupervised learning. The map may be seen to be an equivalent of a domain in a Conceptual Space[1]. The language acquisition process was modeled in a population of simulated agents by using a series of language games, called the observational games. For the experiments, an agent simulation framework was implemented and tested with different parameters. The results of the experiments verify that the agents learn to communicate successfully and a shared lexicon emerges. [6]

Emergence of word features using ICA

We have studied the emergence of linguistic representations through the analysis of words in contexts using the Independent Component Analysis (ICA). The ICA learns features automatically in an unsupervised manner. Several features for a word may exist, and the ICA gives the explicit values of each feature for each word. In our experiments, we have shown that the features coincide with known syntactic and semantic categories. More detailed description of this research is given in the section on Natural Language Processing in this report.

Similarity of emergent representations

According to the connectionist view, mental states consist of activations of neural units in a connectionist network. We consider the similarity of representations that emerge in an unsupervised, self-organization process of neural lattices when exposed to color spectrum stimuli. Self-Organizing Maps (SOM) are trained with color spectrum input, using various vectorial encodings for representation of the input. Further, the SOM is used as a heteroassociative mapping to associate color spectrum with color names. Recall of association between the spectra and colors is assessed, and it is shown that the SOM learns representations for both stimuli and color names, and is able to associate them successfully. The resulting organization is compared through correlation of the activation patterns of the neural maps when responding to color spectrum stimuli. Experiments show that the emerged representations for stimuli are similar with respect to the partitioning-of-activation-space measure almost independently of the encoding used for input representation. This adds new evidence in favour of the usability of the state space semantics.[11]

Self-refreshing SOM as semantic memory

Natural and artificial cognitive systems suffer from forgetting information. However, in natural systems forgetting is typically gradual whereas in artificial systems forgetting is catastrophic. Methods based on rehearsal and pseudorehearsal have been successfully applied in feedforward networks to avoid catastrophic interference. A novel method based on pseudorehearsal for avoiding catastrophic forgetting in the Self-Organizing Map (SOM) is presented. Simulations results show that the use of pseudorehearsal can effectively decrease catastrophic forgetting. [10]

Simulated emotions in a SOM-based agent model

It is assumed that emotions in a cognitive system have a role in interlinking organism's cognition, needs, goals, motivation and final output behavior. For the purpose of emotion modeling, an earlier SOM-based agent simulation model [4] was simplified in many ways. Proposed emotional model might be classified as a cognition appraisal theory inspired by model considering emotions as emergent labels for the evaluation of prototypical situation or events (modal emotions) rather than basic discrete entities achieved by a response program.[14]

Analysis of interprofessional collaboration

The Self-Organizing Map was used to analyze the online collaborative discourses of an interprofessional team of hospital workers in Toronto area engaged in an 18-month reflective practice and continuous learning project. Preliminary results [13] demonstrate unique characteristics of the participant group's interactivity that would otherwise remain unidentified using conventional quantitative methods of discourse analysis. The SOM analysis generated a relational profile of participants' reading and linking activity in an online learning environment that not only captures the emergent dynamics of interprofessional collaboration over time, but also highlights individual differences within and between professional groups.

References

[1] Gärdenfors, P. *Conceptual Spaces.* MIT Press, 2000.

[2] Honkela, T.; Könönen, V.; Pöllä, M.; Simula, O. (eds.) *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05).* Espoo, Finland, June 15-17, 2005. Espoo, Finland 2005. 174 p.

[3] Honkela, T. Von Foerster meets Kohonen - Approaches to Artificial Intelligence, Cognitive Science and Information Systems Development. *Kybernetes*, 34(1/2), 2005.

[4] Honkela, T. and Winter, J. *Simulating Language Learning in Community of Agents Using Self-Organizing Maps.* Helsinki University of Technology, Publications in Computer and Information Science Report, 2003.

[5] Honkela, T.; Hynnä, K.; Lagus, K.; Särelä, J. (eds.) *Adaptive and Statistical Approaches in Conceptual Modeling.* Espoo, Finland: Helsinki University of Technology, 2005. (Publications in Computer and Information Science Technical Report A75).

[6] Lindh-Knuutila, T. *Simulating the Emergence of a Shared Conceptual System in a Multi-Agent Environment.* Master's Thesis. Helsinki University of Technology, Department of Electrical and Communications Engineering, Espoo, Finland. 2005.

[7] Manning, C.D. and Schütze, H. *Statistical Natural Language Processing.* MIT Press, Cambridge, Massachusetts, 1999.

[8] Maturana, H. R. and Varela, F. J. *Autopoiesis and cognition: The realization of the living.* Reidel, Dordrecht, 1980.

[9] Nowé, A., Honkela, T., Könönen, V. and Verbeeck, K. (eds.) *Proceedings of the Workshop W9 on Reinforcement Learning in Nonstationary Environments.* Porto, Portugal, 2005 (in conjunction with the 16th ECML and 9th PKDD, Oct. 3-7, 2005), Portugal 2005, 81 p.

[10] Pöllä, M.; Lindh-Knuutila, T.; Honkela, T. Self-Refreshing SOM as a Semantic Memory Model. *Proc. of AKRR'05, International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning*, Espoo, Finland, June 15-17, 2005, pp. 171-174.

[11] Raitio, J. Vigário, R., Särelä, J. and Honkela, T. Assessing similarity of emergent representations based on unsupervised learning. *Proc. of IJCNN 2004, International Joint Conference on Neural Networks*, Budapest, Hungary, 25-29 July 2004.

[12] Russell, A.; Honkela, T.; Lagus, K.; Pöllä, M. (eds.) *Proceedings of Symposium on Adaptive Models of Knowledge, Language and Cognition (AMKLC'05).* Espoo, Finland, June 15-17, 2005. Espoo, Finland 2005. 61 p.

[13] Russell, A.; Honkela, T. Analysis of interprofessional collaboration in an online learning environment using self-organizing maps. *Proceedings of Symposium on Adaptive Models of Knowledge, Language and Cognition (AMKLC'05)*, Espoo, Finland, June 15-17, pp. 52-57, 2005.

[14] Skripal, P. and Honkela, T. Framework for Modeling Emotions in Communities of Agents. In: H. Hyötyniemi, P. Ala-Siuru and J. Seppänen (eds.), *Life, Cognition and Systems Sciences, Symposium Proceedings of the 11th Finnish Artificial Intelligence Conference*, Finnish Science Center Heureka Vantaa, 1-3 September 2004, pp. 163-172.

[15] Von Foerster, H. (1981). Notes on an epistemology for living things. *Observing Systems.* Intersystems Publications, pp. 257-271. Originally published in 1972 as BCL Report No 9.3., Biological Computer Laboratory, University of Illinois, Urbana.

8.3 Reinforcement learning in multiagent systems

Reinforcement learning methods have attained lots of attention in recent years. Although these methods and procedures were earlier considered to be too ambitious and to lack a firm foundation, they have been established as practical methods for solving, e.g., Markov Decision Processes (MDPs). However, the requirement for reinforcement learning methods to work is that the problem domain in which these methods are applied obeys the Markov property. Basically this means that the next state of a process depends only on the current state, not on the history. In many real-world problems this property is not fully satisfied. However, many reinforcement learning methods can still handle these situations relatively well. Especially, in the case of two or more decision makers in the same system the Markov property does not hold and more advanced methods should be used instead. A powerful tool for handling these highly non-Markov domains is the concept of Markov game. In this project, we have developed efficient learning methods based on the asymmetric learning concept and tested the developed methods with different problem domains, e.g. with pricing applications.

Markov games

With multiple agents in the environment, the fundamental problem of single-agent MDPs is that the approach treats the other agents as a part of the static environment and thus ignores the fact that the decisions of the other agents may influence the state of the environment.

One possible solution is to use competitive multiagent Markov decision processes, i.e. *Markov Games (MGs).* In a MG, the process changes its state according to the action choices of all agents and can thus be seen as a multicontroller MDP. In Fig. 8.1, there is an example of a MG with three states (s_1,s_2,s_3) and two agents. The process changes its state according to probability $P(s_i|s_1, a^1, a^2), i = 2, 3$, where a^1, a^2 are actions selected by the agents 1 and 2.

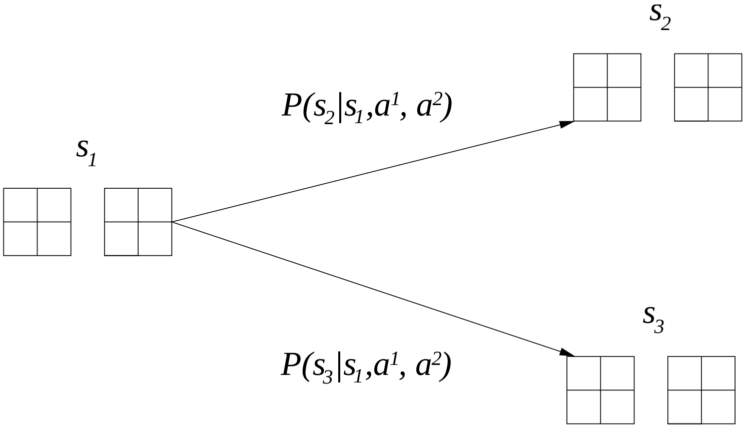


Figure 8.1: An example Markov game with three states.

In single-agent MDPs, it suffices to maximize the utility of the agent in each state. In MGs, however, there are multiple decision makers and more elaborated solution concepts are needed. Game theory provides a reasonable theoretical background for solving this interaction problem. In the single-agent learning, our goal is to find the utility maximizing rule (policy) that stipulates what action to select in each state. Analogously, in a multiagent setting the goal is to find an equilibrium policy between the learning agents.

Practical learning methods

We have concentrated on the case where the state transition probabilities and utility values are not known to the learning agents. Instead, the agents observe their environment and learn from these observations. In general, we use the update rule in the following form:

$$Q^i_{t+1}(s_t, a^1_t, \dots, a^N_t) = (1 - \alpha_t)Q^i_t(s_t, a^1_t, \dots, a^N_t) + \alpha_t[r^i_{t+1} + \gamma f(s_{t+1})], \tag{8.1}$$

where $Q^i_t(s_t, a^1_t, a^2_t)$ is the estimated utility value for the agent i at the time instance t when the system is in the state s_t and agents select actions $a^1_t, \dots, a^N_t$. r^i_{t+1} is the immediate reward for the agent i and γ is the discount factor. f is the function used to evaluate values of the games associated with states. If a symmetric evaluation function is used, i.e. Nash or Correlated Equilibrium function, the update rule is similar for each agent. In the asymmetric case, there is an ordering (some agents make their decisions prior other agents) among learning agents and thus the learning rules are different on different levels of the corresponding agent hierarchy. Further discussion about symmetric learning methods can be found in [1] and [2]. Respectively, fundamental principles and theoretical analysis of the asymmetric model can be found in [3].

Pricing problem in economy

In this section we provide an example of multiagent reinforcement learning. In the problem, there are two competing agents (brokers) that sell identical products and compete against each other on the basis of price. At each time step, one of the brokers decides its new price based on the opponent's, i.e. other broker's, current price. After the prices have been set, the customer either buys a product from the seller or decides not to buy the product at all. The objective of the agents is to maximize their profits.

Tesauro and Kephart modeled [5] the interaction between two brokers as a single-agent reinforcement learning problem in which the goal of the learning agent is to find the pricing strategy that maximizes its long time profits. Additionally, reinforcement learning aids the agents to prevent "price wars", i.e. repeated price undercutting among the brokers. As a consequence of a price war, the prices would become very small and the overall profits would also be small. Tesauro and Kephart reported very good performance of the approach when one of the brokers keeps its pricing strategy fixed. However, if both brokers try to learn simultaneously, the Markov property assumed in the theory of MDPs does not hold any more and the learning system encounters serious convergence problems. For solving these convergence problems we have modeled the system as a Markov game. In the example depicted in Fig. 8.2, cumulative profits that are averages of 1000 test runs each containing 10 pricing decision for both brokers are plotted against the planning depth (discount factor γ). In this simple example all prices were between $[0, 1]$ and the customer bought the product from the broker with the lowest price. More discussion on the pricing problem can be found in [4].

References

[1] A. Greenwald and K. Hall. Correlated-Q learning. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML-2003)*, pages 242–249, Washington, DC, 2003.

[2] J. Hu and M. P. Wellman. Nash Q-learning for general-sum stochastic games. *Journal of Machine Learning Research*, 4:1039–1069, 2003.

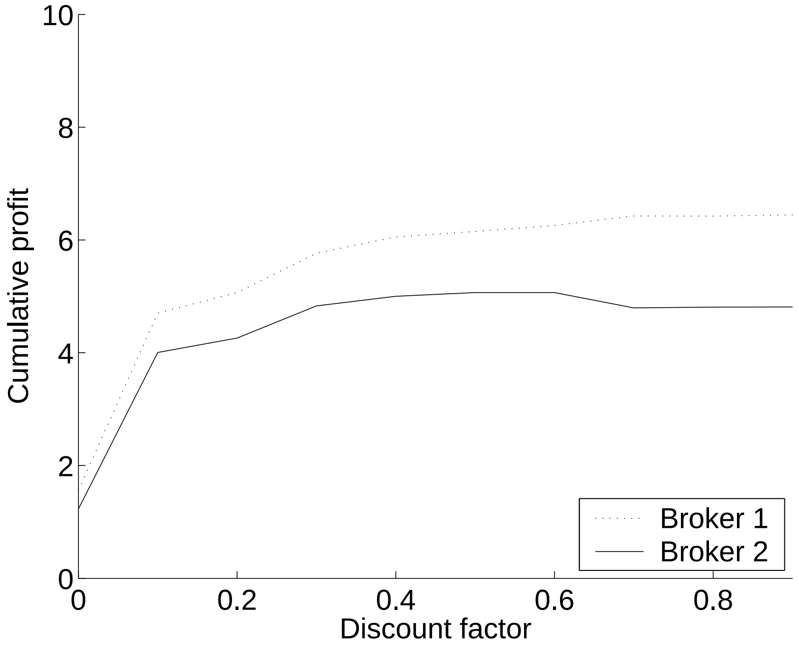


Figure 8.2: Averaged profits in the pricing example. All data points are averages of 1000 test runs each containing 10 pricing decisions for both agents.

[3] V. J. Könönen. Asymmetric multiagent reinforcement learning. *Web Intelligence and Agent Systems: An International Journal (WIAS)*, 2(2):105–121, 2004.

[4] V. J. Könönen. Dynamic pricing based on asymmetric multiagent reinforcement learning. *International Journal of Intelligent Systems*, 21(1):73–98, 2006.

[5] G. Tesauro and J. O. Kephart. Pricing in agent economies using multi-agent Q-learning. In *Proceedings of the Workshop on Game Theoretic and Decision Theoretic Agents (GTDT'99)*, pages 71–86, London, UK, 1999.

Chapter 9

Speech recognition

Mikko Kurimo, Panu Somervuo, Kalle Palomäki, Vesa Siivola, Teemu Hirsimäki, Janne Pylkkönen, Simo Broman, Ville Turunen, Sami Virpioja

9.1 The speech recognition tasks and systems

This chapter is divided into four categories that describe our research activities in: **1.Acoustic modeling, 2.Language modeling, 3.Large vocabulary decoders, and 4.Speech retrieval**. The division is natural both because it covers four of the major subfields in speech recognition research and because it describes the main components of a typical large vocabulary continuous speech recognition (LVCSR) system (Figure 9.1). The acoustic models produce the probabilities of different phonemes, the language models take into account the co-occurrence probabilities of different words or morphemes, the decoder joins these two streams of information into recognition hypothesis, and the retrieval engine utilizes these outputs to represent the speech in a convenient form for searching and browsing. Thus, all our research topics focus on the same framework and can be integrated into a single working LVCSR system.

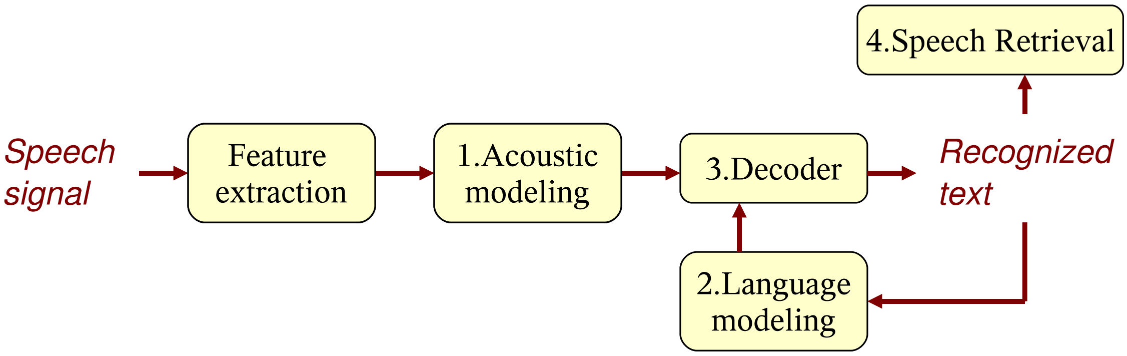


Figure 9.1: The main components of the LVCSR system.

Our goal in LVCSR research has for several years been to develop new machine learning algorithms for each of the subfields and build a complete state-of-art recognizer to evaluate the new methods and their impact. Originally, the recognizer was constructed for fluent and planned speech such as Finnish newsreading, where language models covering a very large vocabulary are required. Besides newsreading, other example tasks are political and academic speeches and other radio and television broadcasts where the language used is near the written style. Sofar, we have not seriously attempted to recognize spontaneous conversations, because enough Finnish training texts for learning the corresponding style do not exist. Our main training corpus for language modeling is the Finnish Language Bank at CSC. For acoustic modeling we use voice books, Finnish Broadcast Corpus at CSC and the SPEECON corpus.

In addition to the recognition of Finnish, we have performed experiments in English, Turkish and Estonian. To make this possible we have established research relations to different top speech groups in Europe and U.S., e.g. University of Colorado, International Computer Science Institute, IDIAP, University of Edinburgh, University of Sheffield, Bogazici University, and Tallinn University of Technology. The forms of collaboration have included researcher exchanges, special courses, workshops and joint research projects. We have also participated in several top international and national research projects funded by EU, Academy of Finland, Tekes, and our industrial partners. In the close collaboration with our Natural Language Processing group 10 we are also organizing an international competition called Morphochallenge to evaluate the best unsupervised segmentation algorithms for words into morphemes for LVCSR and language modeling in different languages. This challenge project is funded by EU's PASCAL network.

9.2 Acoustic modeling

Phoneme modeling and speaker adaptation

Acoustic modeling in automatic speech recognition (ASR) means building statistical models for some meaningful speech units based on the feature vectors computed from speech. In most systems the speech signal is first chunked into overlapping 20-30 ms time windows at every 10 ms and the spectral representation is computed from each frame. A commonly used feature vector consists of mel-frequency cepstral coefficients (MFCC) which are the result of the discrete cosine transform (DCT) applied to the logarithmic mel-scaled filter bank energies. Local temporal dynamics can be captured by concatenating the first and second order delta features (time differences) to the basic feature vector.

After the feature extraction the feature sequence is typically modeled using hidden Markov models (HMM). In basic form each phoneme is modeled by a separate HMM, where the emission distributions of the HMM states are Gaussian mixtures. An example is shown in Figure 9.2. In practice, however, we need to take the phoneme context into account, so that for each phoneme there are separate HMMs for various phoneme contexts. Even though the Gaussian mixture components are restricted to have only diagonal covariance matrices, the number of parameters with such a complex acoustic model in a typical state-of-the-art ASR system is very high, in order of millions of parameters. This gives emphasis to proper complexity control, so that we get the most out of the available training data.

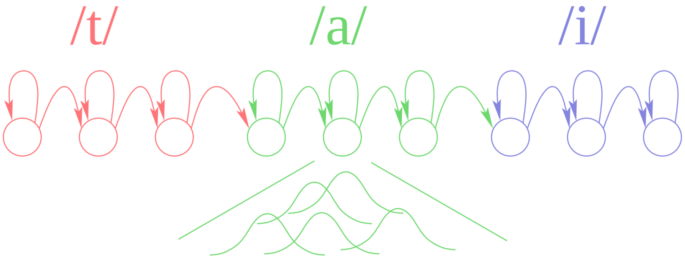


Figure 9.2: Each phoneme is modeled with a hidden Markov model, usually consisting of three states. The state distributions are modeled by Gaussian mixture models.

The problem of selecting the optimal model complexity is a difficult one, but it can be avoided by using some other model training criterion instead of the usual maximum likelihood (ML) principle. In [1] two more advanced training principles, maximum a posteriori (MAP) and variational Bayesian (VB), were compared against the ML principle. These two methods can avoid overfitting in case of too complex a model, which is the major drawback in ML training. This was also validated experimentally, where the speech recognition performance started to degrade with ML trained models when the number of parameters was increased, whereas MAP and VB trained models continued to work well. The VB principle can also be used to select the proper model complexity in respect to the training data, without using auxiliary data.

Hidden Markov models have several drawbacks with respect to speech modeling. One of those is the modeling of the durations of speech segments. Standard HMMs allow only minimal modeling of duration variations, although in some languages (e.g. in Finnish) the durations can be the main cues in discriminating between certain phonemes. To better take the durations into account we experimented in [2] several extensions to standard HMMs which allow more precise models for the segmental durations. It was found out that already a relatively simple duration model was enough to improve the speech recognition results.

In the past most of our speech recognition experiments have been carried out with

Method	Word error rate (%)	Phoneme error rate (%)
Baseline	30.5	9.8
VTLN	29.3	9.1
cMLLR	25.3	7.3
SAT/cMLLR	24.2	6.8

Table 9.1: Speech recognition results for several adaptation methods: No adaptation (Baseline), Vocal Tract Length Normalization (VTLN), Constrained Maximum Likelihood Linear Regression (cMLLR) and Speaker Adaptive Training with cMLLR (SAT/cMLLR).

speaker dependent models, meaning the acoustic models have been trained specifically for one person. Recently we have been able to move to more demanding speaker independent experiments, which is also more realistic in view of many applications. The lack of speaker dependency adds further demands for the acoustic models. We have therefore implemented several adaptation techniques to our speech recognizer, and tested their effectivity with our speech data. Some results are shown in Table 9.1.

Recognition of reverberant speech

In the acoustic modeling for large vocabulary continuous speech recognition mostly speech in relatively noise free condition was concentrated (see Sect. 9.2). In the field of noise robust speech recognition, we have been developing techniques to recognition of reverberant speech jointly with the University of Sheffield [4]. Our technique is based to missing data approach [5], in which a conventional Gaussian mixture model classifier is adapted to allow different treatments of reliable and unreliable regions of speech. In our approach the regions of speech spectrum, which are either relatively clean or badly contaminated by reverberation are indexed and used to construct a time frequency mask to the missing data speech recognizer. Masks are produced by applying modulation filtering to detect strong speech regions not contaminated by reverberation (see Fig. 9.3). Furthermore, we were able to improve the performance slightly by combining the missing data recognizer to a conventional recognizer using cepstral features. More information about techniques to handle reverberation in the auditory scene analysis as well as in speech recognition can be read from our recent review article [3].

References

[1] P. Somervuo: Comparison of ML, MAP, and VB based acoustic models in large vocabulary speech recognition In *Proceedings of the 8th International Conference on Spoken Language Processing* (Interspeech 2004), October 4–8, 2004, Jehu Island, Korea, pp. 701–704.

[2] J. Pylkkönen and M. Kurimo: Duration Modeling Techniques for Continuous Speech Recognition In *Proceedings of the 8th International Conference on Spoken Language Processing* (Interspeech 2004), October 4–8, 2004, Jeju Island, Korea, pp. 385–388.

[3] G. J. Brown and K. J. Palomäki Reverberation, in *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, eds. by DeLiang Wang and Guy J. Brown, Wiley/IEEE Press, to appear in June 2006.

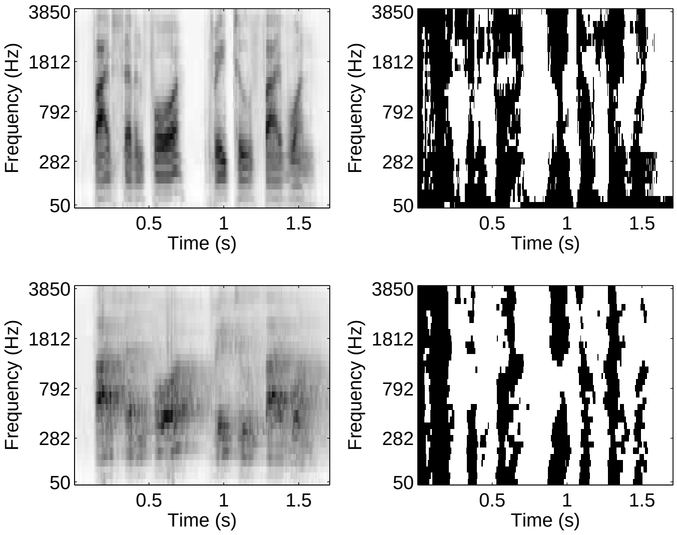


Figure 9.3: Auditory spectrograms (left panels) and missing data masks (right panels). The spectrogram for a "clean" unreverberated (top-left) and a highly reverberant (bottom left) speech utterance are shown. The right panels show the corresponding missing data masks for the reverberant utterance. Firstly, an "oracle mask" based on prior knowledge (top-right) of the reverberated regions shows how an (nearly) ideal mask should look like. Secondly, a mask produced using our model (with no prior knowledge) is shown. Black and white regions indicate reliable and unreliable regions, respectively.

[4] K. J. Palomäki, G. J. Brown and J. Barker, Recognition of reverberant speech using full cepstral features and spectral missing data Accepted for publication in *Proc. ICASSP 2006.*

[5] M.P. Cooke, P. Green, L. Josifovski, and A. Vizinho, Robust automatic speech recognition with missing and unreliable acoustic data,*Speech Comm.*, vol. 34, pp. 267 285, 2001.

9.3 Language modeling

Splitting words into fragments

For Finnish, estimating the language model probabilities for words is difficult since there is a vast number of different word forms. For example, a single verb has theoretically thousands of inflected word forms.

The natural way to attack the problem is to split words into smaller fragments and build the language models on the fragments instead of whole words. Since it is not obvious how the words should be split, we have studied what kind of word fragments are optimal for speech recognition systems. Experiments in Finnish recognition tasks indicate that an unsupervised data-driven splitting algorithm called Morfessor (see Section 10.1) can produce word fragments that work even better in speech recognition than morphemes based on Finnish grammar [1, 2].

Since the Morfessor algorithm is language independent, it can also be applied to speech recognition of other languages. Experiments in Turkish and Estonian recognition tasks confirm the result that models based the Morfessor algorithm improve recognition accuracy.

A growing method for constructing an n-gram model

The length of the word history used by the n-gram model is traditionally set to a fixed n. For $n > 3$ this often leads to prohibitively big models. We have developed an algorithm based on the Minimum Description Length principle [3], which learns a suitable word history length for each case [4]. The factors affecting the choice of histories are: 1) Does the model get much better if we use a longer word history for modeling an n-gram? and 2) Do we have enough data to estimate the probabilities for the longer history? This method can make considerably smaller n-gram models which equal modeling power of the fixed n models.

A related method is the pruning of n-gram models, for example entropy based pruning [5]. The benefits of our approach compared to pruning methods are that at no time we need to store the full model. This allows us to train very high order models. Our experiments show, that the growing method seems to outperform the entropy based pruning in practically all experiments [6]. For example in Finnish speech recognition experiments, the growing method gives at least 15% lower word error rate (relative) for reasonable n-gram model sizes, when both methods use equal model size. The n-gram models can be efficiently stored in a tree structure (Fig. 9.4).

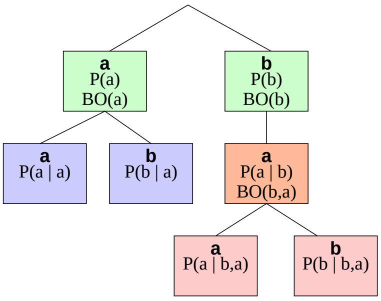


Figure 9.4: The tree structure for storing an n-gram model.

Combining the growing method and clustering

In addition to the pruning of the n-gram models, a common way to decrease the size of the n-gram models is clustering of model units or sequences of them. In a similar manner that the MDL principle can be used to choose a suitable length of n-gram history for each case, it can be used to insert n-gram histories that give similar predictions into same equivalence classes [7]. Compared to the baseline of the growing method, for a model of an equal size, some accuracy may be lost, but substantially more n-grams can be included into the model.

In order to make the clustering computationally fast, the number of different model units cannot be very large. Suitable small lexicons are easy to construct for any language with the Morfessor algorithm (Section 10.1). Preliminary experiments show that with some optimizations, even extensive searches for the nearest history clusters are possible. This differs from e.g. one related method [8], where only nearby parts of the tree structure are searched for similar prediction distributions.

Combining methods for language models

In many task the best language modeling results have been achieved when different language models have been used together [9]. Several combination methods have been presented in the literature, but a thorough investigation of the methods has not been done.

In [10, 11], the combination methods that have been used with language models are studied. Also, a new approach based on likelihood density function estimation using histograms is presented. In addition to theoretical consideration, four combining methods for four language models are evaluated in speech recognition experiments and word prediction experiments using Finnish news articles.

In the perplexity experiments, all combining methods produced statistically significant improvement compared to the 4-gram model that worked as a baseline. The best result, 46 % improvement to the 4-gram model, was achieved when combining several language models together by using the new bin estimation method. In the speech recognition experiments, 4 % reduction to the word error and 7 % reduction to the phoneme error was achieved.

References

[1] T. Hirsimäki, M. Creutz, V. Siivola, M. Kurimo, S. Virpioja, J. Pylkkönen, Unlimited vocabulary speech recognition with morph language models applied to Finnish. *Computer Speech & Language*, accepted for publication in 2005.

[2] T. Hirsimäki, M. Creutz, V. Siivola and M. Kurimo: Morphologically Motivated Language Models in Speech Recognition, In *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning* (AKRR05), June 15-17, 2005, Espoo, Finland, pp. 121-126.

[3] Rissanen, J., 1994. *Language Computation.* American Mathematical Society, Ch. Language Acquisition in the MDL Framework.

[4] V. Siivola, "Building compact language models incrementally," in *Proceedings of Second Baltic Conference on Human Language Technologies*, 2005, pp. 183–188.

[5] Stolcke, A., 1998. Entropy-based pruning of backoff language models. *In Proc. DARPA Broadcast News Transcription and Understanding Workshop.* pp. 270–274.

[6] Siivola, V., Pellom, B., 2005. Growing an n-gram model. *In Proc. Interspeech 2005.* pp. 1309–1312

[7] Virpioja, S., 2005. New methods for statistical natural language modeling. Master's thesis, Department of Computer Science and Engineering, Helsinki University of Technology.

[8] Siu, M., Ostendorf, M., 2000. Variable n-grams and extensions for conversational speech language modeling. *IEEE Transactions on Speech and Audio Processing*, 8(1):63–75.

[9] Kurimo, M., Zhou, B., Huang, R., Hansen, J.H.L., 2004. Language modeling structures in audio transcription for retrieval of historical speeches. *In Proc. EUSIPCO 2004.*

[10] Broman, S., 2005. Combining methods for language models in speech recognition. Master's thesis, Helsinki University of Technology.

[11] Broman, S., Kurimo, M., 2005. Methods for combining language models in speech recognition. *In Proc. Interspeech 2005.* pp. 1317–1320

9.4 Large vocabulary decoder

The goal of the speech recognition is to find the word sequence that is the most probable one given the acoustic model, language model and the observed speech. Because the number of possible word sequences is extremely high, the search is performed incrementally in time, and improbable hypotheses are pruned as early as possible. The module responsible for performing this search is called the decoder.

During the recent years, we have actively developed a decoder for very large vocabularies. In order to take the acoustic dependencies better into account, the stack based decoder has been replaced by an efficient time-synchronous token-pass decoder [1]. The efficiency is derived from a compact search network which can utilize the redundancies in the acoustic models. The decoder is able to model correctly also the context dependent phonemes which occur across lexical units. Compared to our previous decoder, this results in a 24% relative improvement of the phoneme error rate.

The efficiency of decoders continues to be an important issue in speech recognition, as more and more complex models of acoustics and language are used to obtain the possible best recognition accuracy. Some new ways to restrict the search space without affecting the recognition accuracy too much were developed in [2]. These so called pruning criteria use different information available during the search to discard those path hypotheses which no longer seem feasible. The research also resulted in a method with which we can avoid hand tuning the numerous parameters affecting the efficiency/accuracy tradeoff in the decoding process.

References

[1] J. Pylkkönen: An Efficient One-pass Decoder for Finnish Large Vocabulary Continuous Speech Recognition, In *Proceedings of the 2nd Baltic Conference on Human Language Technologies* (HLT'2005), April 4–5, 2005, Tallinn, Estonia, pp. 167–172.

[2] J. Pylkkönen: New Pruning Criteria for Efficient Decoding, In *Proceedings of the 9th European Conference on Speech Communication and Technology* (Interspeech 2005), September 4–8, 2005, Lisboa, Portugal, pp. 581–584.

9.5 Spoken document retrieval

Speech retrieval and indexing

One important application of automatic speech recognition is spoken document retrieval which means the task of finding interesting segments from recorded speech. Large amount of information is produced in spoken form, for example radio and TV broadcasts, and there is a need for tools that can be used to search this data. Spoken document retrieval systems combine speech recognition and information retrieval technologies. The special properties of the Finnish language, such as the large number of inflected word forms, affect both of these parts and methods developed for other languages cannot be used as such. Our research is focused on retrieval of Finnish speech, but the methods are hoped to work also on other languages with similar properties.

Previously, word-based and phone-based approaches have been used. The former suffers from limited vocabulary and the latter from high error rates. In [1, 2], we presented a baseline retrieval system for Finnish that uses the speech recognizer based on morpheme-like subword units. The recognizer can achieve low error rates while providing a potentially unlimited vocabulary. Retrieval performance of spoken news was found to be close to that of the human reference transcripts. The morpheme like units were found to work well also as index terms, providing equal performance to base formed words.

Recognition errors degrade retrieval performance, but there are measures that can be used to reduce their effect. For example, the recognizer can be modified to include alternative recognition results in the transcripts, or queries can be expanded by adding relevant words from a parallel text corpus. Query expansion was found to bring the level of performance to the same as text document retrieval, even for transcriptions with relatively high error rates. [3, 4]

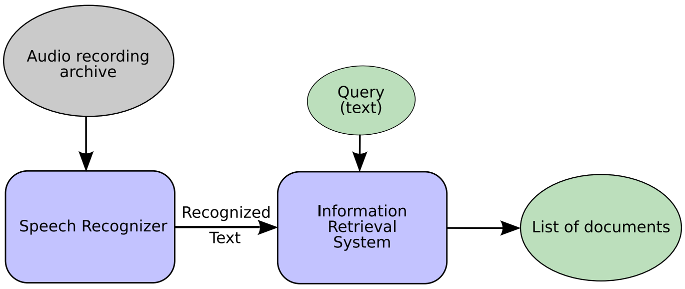


Figure 9.5: Overview of a spoken document retrieval system.

Speech segmentation

The development of automatic segmentation methods of speech and audio is increasingly important to allow automatic handling of growing archives of spoken audio, e.g. recorded meetings, radio or television programs. Audio material can be segmented based on various levels of description. On metadata level audio can be classified e.g. to speech vs. multiple classes of non-speech. Furthermore, segments containing only speech can be classified based on gender, speaker identity and, finally, into subunits of speech, such as, sentences, words or phonemes. Segmentation can be performed either in a supervised or unsupervised manner. In the supervised segmentation, the task is to align temporal structure of speech to the existing transcription. In the unsupervised segmentation the transcript does not exist, and the recognizer classifies the segments freely.

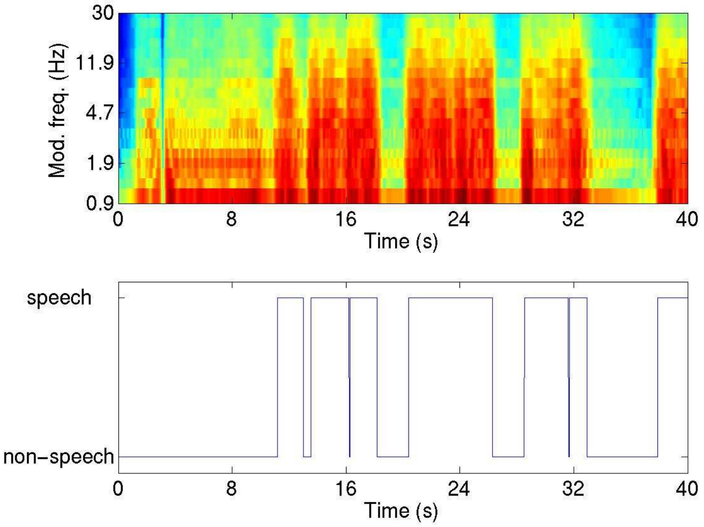


Figure 9.6: Speech modulation spectrum (top) against target classification (bottom).

The group's speech recognition tools [1] were applied to supervised phoneme level segmentation to align existing transcriptions to the corresponding speech audio. This line of research was extended in a student project to speech that is only partially transcribed. Moreover, practical speech recognition tasks have prompted researchers in the group to develop unsupervised techniques [5] to speaker segmentation. These methods were also used to address the needs of other speech researchers in the Helsinki University, Tampere University of Technology and University of Turku.

A new research project in speech segmentation in the metadata level was initiated during April 2005. In this project, we are developing techniques to extract speech segments from audio stream, and speech segments based on gender in an unsupervised manner. The project was started with development of feature techniques using a common Gaussian mixture model classifier. In our new approach, we have applied two types of feature presentations of speech, first, which depicts short-term ($\approx$ 16 ms) speech spectrum and the second that describes long term temporal modulations ($\approx$ 1 s) of speech applying a computation of modulation spectrum (see Fig. 9.6).

References

[1] M. Kurimo, V. Turunen and I. Ekman: An Evaluation of a Spoken Document Retrieval Baseline System in Finnish, *ICSLP*, 2004.

[2] M. Kurimo, V. Turunen and I. Ekman: Speech Transcription and Spoken Document Retrieval in Finnish, *Machine Learning for Multimodal Interaction* Revised Selected Papers of the MLMI 2004 workshop. Lecture Notes in Computer Science, Vol. 3361, pages 253–262, 2005

[3] M. Kurimo and V. Turunen: To recover from speech recognition errors in spoken document retrieval, *Interspeech 2005*, pages 605–608, 2005.

[4] V. Turunen, Spoken document retrieval in Finnish based on morpheme-like subword units, M.S. thesis, Helsinki University of Technology, Espoo, Finland, 2005.

[5] L. Wilcox, F. Chen, D. Kimber and V. Balsubramaman Segmentation of speech using speaker identification. *Proc. ICASSP 1994*, pp. I-161-I-164.

Chapter 10

Natural language processing

Krista Lagus, Timo Honkela, Mathias Creutz, Mikko Kurimo, Sami Virpioja, Jaakko Väyrynen, Oskar Kohonen, and Krister Lindén

10.1 Unsupervised segmentation of words into morphs

In the theory of linguistic morphology, morphemes are considered to be the smallest meaning-bearing elements of language, and they can be defined in a language-independent manner. It seems that even approximative automated morphological analysis is beneficial for many natural language applications dealing with large vocabularies, such as speech recognition and machine translation. Many existing applications make use of *words* as vocabulary units. However, for some languages, e.g., Finnish and Turkish, this leads to very sparse data, as the number of possible word forms is very high. Figure 10.1 shows the very different rates at which the vocabulary grows in Finnish and English text corpora of the same size. The number of different unique word forms in the Finnish corpus is considerably higher than in the English one.

We have developed *Morfessor*, a language-independent, data-driven method for the unsupervised segmentation of words into morpheme-like units. There are different versions of Morfessor, which correspond to consecutive steps in the development of the model [1, 2, 3, 4]. All versions can be seen as instances of a general model, as described in [5].

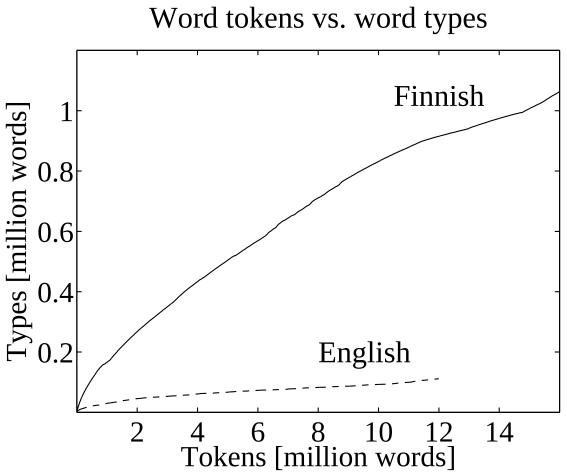


Figure 10.1 The number of different word forms (types) encountered in growing portions of running text (tokens) of Finnish and English.

The general idea behind the Morfessor model is to discover as compact a description of the data as possible. Substrings occurring frequently enough in several different word forms are proposed as *morphs* and the words are then represented as a concatenation of morphs, e.g., “hand, hand+s, left+hand+ed, hand+ful”.

An optimal balance is sought between compactness of the *morph lexicon* versus the compactness of the representation of the *corpus*. The morph lexicon is a list of all distinct morphs (e.g., “hand, s, left, ed, ful”) together with some stored properties of these morphs. The representation of the corpus can be seen as a sequence of pointers to entries in the morph lexicon; e.g. the word “lefthanded” is represented as three pointers to morphs in the lexicon.

Among others, de Marcken [6], Brent [7], and Goldsmith [8] have shown that the above type of model produces segmentations that resemble linguistic morpheme segmentations, when formulated mathematically in a probabilistic framework or equivalently using the Minimum Description Length (MDL) principle [9].

A shortcoming of previous splitting methods is that they either do not model *context-dependency* or they *limit the number of splits* per word to two or three. For instance, failure to incorporate context-dependency in the model may produce splits like “s+wing, ed+ward, s+urge+on” on English data, since the morphs “-s” and “-ed” are frequently occurring suffixes in the English language, but the algorithm does not make this distinction and thus suggests them in word-initial position as prefixes. By limiting the number of allowed segments per word the search task is alleviated and context-dependency can be modeled. However, this makes it impossible to correctly segment compound words with several affixes (pre- or suffixes), such as the Finnish word “aka+n+kanto+kiso+i+ssa”

aarre + **kammio** + *i* + *ssa*, **aarre** + **kammio** + *nsa*, **bahama** + **saar** + *et*,
bahama + **saari** + *lla*, **bahama** + **saar** + *ten*, **edes** + **autta** + *isi* + *vat*,
edes + **autta** + *ma* + *ssa*, nais + **auto** + *ili* + *ja* + *a*, pää + **aihe** + *e* + *sta*,
pää + **aihe** + *i* + *sta*, **pää** + *hän*, taka + **penkki** + *lä* + *in* + *en*, **voi** + *mme* + *ko*

abandon + *ed*, **abandon** + *ing*, **abandon** + *ment*, **beauti** + *ful*,
beauty + *'s*, **calculat** + *ed*, **calculat** + *ion* + *s*, **express** + *ion* + *ist*,
micro + **organ** + *ism* + *s*, **long** + **fellow** + *'s*, **master** + **piece** + *s*,
near + *ly*, **photograph** + *er* + *s*, **phrase** + *d*, un + **expect** + *ed* + *ly*

ansvar + *ade*, **ansvar** + *ig*, **ansvar** + *iga*, **ansvar** + *s* + för + **säkring** + *ar*,
blixt + ned + **slag**, **dröm** + *de*, **dröm** + *des*, **drömma** + *nde*, in + **lopp** + *et* + *s*,
in + **lägg** + *n* + *ing* + *ar*, **målar** + *e*, **målar** + **yrke** + *t* + *s*, o + ut + **nyttja** + *t*,
poli + *s* + **förening** + *ar* + *na* + *s*, **trafik** + **säker** + *het*, över + **fyll** + *d* + *a*

Figure 10.2: Examples of segmentations learned from data sets of Finnish, English, and Swedish text. Suggested prefixes are underlined, stems are rendered in **boldface**, and suffixes are *slanted*.

(transl. "in the wife-carrying contests").

We have focused our efforts on developing a segmentation model that incorporates context-dependency without restricting the number of allowed segments per word. This has resulted in two model variants [3, 4]. The former is based on Maximum Likelihood (ML) optimization, in combination with some heuristics. The latter constitutes an attempt to more elegant model formulation, within the Maximum a Posteriori (MAP) framework.

Evaluation

Morfessor has been evaluated in two complementary ways: directly by comparing to linguistic morpheme segmentations of Finnish and English words, and indirectly as a component of a large (or virtually unlimited) vocabulary Finnish speech recognition system. In both cases, Morfessor outperforms state-of-the-art solutions. The speech recognition experiments are described in Section 9.3.

In order to carry out the direct evaluation, linguistic reference segmentations needed to be produced as part of the current project, since no available resources were applicable as such. This work has resulted in a morphological "gold standard", called *Hutmegs* (Helsinki University of Technology Morphological Evaluation Gold Standard) [10, 11]. Hutmegs contains analyses for 1.4 million Finnish and 120 000 English word forms, which have been produced by further processing the contents of the Finnish Two-Level Morphological Analyzer from Lingsoft, Inc. and the English CELEX database from the Linguistic Data Consortium (LDC). Hutmegs is publicly available for research; inexpensive one-time license fees need to be paid to Lingsoft and the LDC, for access to the Finnish and English analyses, respectively.

When the latest context-sensitive Morfessor versions [3, 4] are evaluated against the Hutmegs gold standard, they clearly outperform a frequently used benchmark algorithm [8] on Finnish data, and perform as well or better than the benchmark on English data (depending on the size of the data sets used).

Some sample segmentations of Finnish, English, as well as Swedish words, are shown in Figure 10.2. These include correctly segmented words, where each boundary coincides with a linguistic morpheme boundary (e.g., "aarre+kammio+i+ssa, edes+autta+isi+vat, abandon+ed, long+fellow+'s, in+lopp+et+s"). In addition, some words are over-segmented, with boundaries inserted at incorrect locations (e.g., “in+lägg+n+ing+ar” instead of “in+lägg+ning+ar”), as well as under-segmented words, where some boundary is missing (e.g., “bahama+saari+lla” instead of “bahama+saar+i+lla”).

In addition to segmenting words, Morfessor suggests likely grammatical categories for the segments. Each morph is tagged as a prefix, stem, or suffix. Sometimes the morph categories can resolve the semantic ambiguity of a morph, e.g., Finnish “pää”. In Figure 10.2, “pää” has been tagged as a stem in the word “pää+hän” (“in [the] *head*”), whereas it functions as a prefix in “pää+aihe+e+sta” (“about [the] *main* topic”).

Demonstration and software

There is an online demonstration of Morfessor on the Internet: `http://www.cis.hut.fi/projects/morpho/`. Currently, the demo supports three languages: Finnish, English, and Swedish. Those interested in larger-scale experiments can download the Morfessor program and train models using their own data sets. The software is described in [12]. Within a period of ten months (April 2005 – January 2006) a monthly average of 18 downloads of the program has been registered.

Further applications

Outside the scope of the current project, Morfessor has been used successfully in the recognition of Turkish [13] as well as Estonian speech. Hagen and Pellom [14] apply Morfessor in English speech recognition intended for oral reading tracking within an interactive reading tutor program for children. Morfessor has also been used in Finnish information retrieval, both in the retrieval of text [15] and spoken documents [16] (Sec. 9.5). Furthermore, in a number of works on language modeling, the segments discovered by Morfessor constitute the basic vocabulary [17, 18, 19] (Section 9.3). Kumlander [20] has analyzed the word splits obtained when running Morfessor on stories told by Finnish children.

References

[1] Mathias Creutz and Krista Lagus. Unsupervised discovery of morphemes. In *Proc. Workshop on Morphological and Phonological Learning of ACL’02*, pages 21–30, Philadelphia, Pennsylvania, USA, 2002.

[2] Mathias Creutz. Unsupervised segmentation of words using prior distributions of morph length and frequency. In *Proc. ACL’03*, pages 280–287, Sapporo, Japan, 2003.

[3] Mathias Creutz and Krista Lagus. Induction of a simple morphology for highly-inflecting languages. In *Proc. 7th Meeting of the ACL Special Interest Group in Computational Phonology (SIGPHON)*, pages 43–51, Barcelona, July 2004.

[4] Mathias Creutz and Krista Lagus. Inducing the morphological lexicon of a natural language from unannotated text. In *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR’05)*, 2005.

[5] Mathias Creutz and Krista Lagus. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing*, 2006. (Accepted for publication).

[6] C. G. de Marcken. *Unsupervised Language Acquisition.* PhD thesis, MIT, 1996.

[7] M. R. Brent. An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, 34:71–105, 1999.

[8] John Goldsmith. Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27(2):153–198, 2001.

[9] Jorma Rissanen. *Stochastic Complexity in Statistical Inquiry*, volume 15. World Scientific Series in Computer Science, Singapore, 1989.

[10] Mathias Creutz and Krister Lindén. Morpheme segmentation gold standards for Finnish and English. Technical Report A77, Publications in Computer and Information Science, Helsinki University of Technology, 2004.

[11] Mathias Creutz, Krista Lagus, Krister Lindén, and Sami Virpioja. Morfessor and Hutmegs: Unsupervised morpheme segmentation for highly-inflecting and compounding languages. In *Proceedings of the Second Baltic Conference on Human Language Technologies*, pages 107–112, Tallinn, Estonia, 4 – 5 April 2005.

[12] Mathias Creutz and Krista Lagus. Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0. Technical Report A81, Publications in Computer and Information Science, Helsinki University of Technology, 2005.

[13] K. Hacioglu, B. Pellom, T. Ciloglu, O. Ozturk, M. Kurimo, and M. Creutz. On lexicon creation for Turkish LVCSR. In *Proc. Eurospeech'03*, pages 1165–1168, Geneva, Switzerland, 2003.

[14] Andreas Hagen and Bryan Pellom. Data driven subword unit modeling for speech recognition and its application to interactive reading tutors. In *Proceedings of INTERSPEECH 2005*, pages 2757–2760, Lisbon, Portugal, September, 4–8 2005.

[15] Sam Engström. Information retrieval using unsupervisedly segmented morphemes. Special assignment, Laboratory of Computer and Information Science, Helsinki University of Technology, July 2005.

[16] Mikko Kurimo and Ville Turunen. To recover from speech recognition errors in spoken document retrieval. In *Proceedings of Interspeech 2005*, pages 605–608, Lisbon, Portugal, September 2005.

[17] Vesa Siivola and Bryan L. Pellom. Growing an n-gram language model. In *Proceedings of Interspeech 2005*, Lisbon, Portugal, September 2005.

[18] Simo Broman and Mikko Kurimo. Methods for combining language models in speech recognition. In *Proceedings of Interspeech 2005*, pages 1317–1320, Lisbon, Portugal, September 2005.

[19] Sami Virpioja. New methods for statistical language modeling. Master's thesis, Helsinki University of Technology, Department of Computer Science and Engineering, Laboratory of Computer and Information Science, 2005.

[20] Mikaela Kumlander. Forthcoming master's thesis. Master's thesis, University of Helsinki, Department of General Linguistics, 2005.

10.2 Word sense disambiguation using document maps

Krister Lindén, Krista Lagus

A single word may have several senses or meanings, for example "was *heading* south/the newspaper *heading* is", or "Church" as an institution versus "church" as a building. Word sense disambiguation automatically determines the appropriate senses of a particular word in context. It is an important and difficult problem with many practical consequences for language-technology applications in information retrieval, document classification, machine translation, spelling correction, parsing, and speech synthesis as well as speech recognition. For a textbook introduction, see [5]. In particular, Yarowsky [6] noted that words tend to keep the same sense during a discourse.

In [2] we introduce a method called THESSOM for word sense disambiguation that uses an existing topical document map, in this case a map of nearly 7 million patent abstracts, created with the WEBSOM method (see [1]). The method uses the document map as a representation of the semantic space of word contexts. The assumption is that similar meanings of a word have similar contexts, which are located in the same area on the self-organized document map. The results confirm this assumption. In this method, the existing general-purpose document map is calibrated, i.e., marked with correct senses, using a subset of data where the ambiguous words have been sense-tagged. The sense-calibrated map can then be utilized as a word sense classifier, for determining a probable correct sense for an ambiguous sample word in context. The data flow of the training and testing procedure is shown in Figure 10.3.

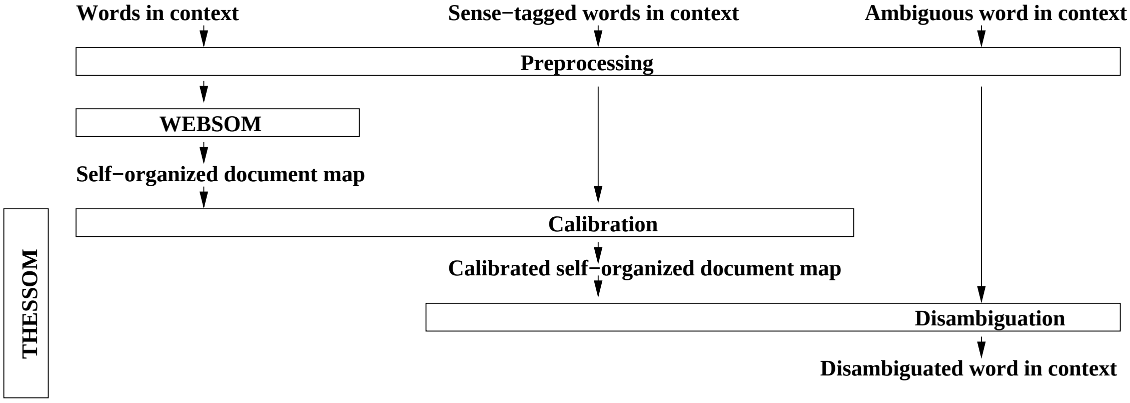

Words in context
Sense−tagged words in context
Ambiguous word in context
Preprocessing
WEBSOM
Self−organized document map
Calibration
Calibrated self−organized document map
Disambiguation
Disambiguated word in context
THESSOM

Figure 10.3: Data flow of word sense disambiguation with self-organized document maps

Results on the SENSEVAL-2 corpus (from a word sense disambiguation contest) indicate that the proposed method is statistically significantly better than the baselines, and performs on an average level when compared to the total of supervised methods in the competition. The benefit of the proposed method is that a single general purpose representation of the semantic space can be used for all words and their word senses.

In [3], instead of utilizing one general-purpose document map and merely calibrating (marking) it with particular sense locations, an individual document map is created for each ambiguous word from the training material (short contexts) for that word. Moreover, advanced linguistic analysis was performed using a dependency grammar parser to produce additional features for the document vectors. The training material consisted of a total of 8611 contexts for the 73 ambiguous words, i.e., on the average 118 contexts per word. As a result, 73 maps were generated, one for each ambiguous word.

In [4], we evaluate the efficacy of various features for word sense disambiguation with THESSOM. We conclude that the syntactic features are the most important for word sense disambiguation and should be chosen if a single feature type needs to be selected. However, their feature space is sparse, so the features based on the base forms of words function as a kind of back-off model with a small but statistically significant improvement to the overall dismabiguation performance of THESSOM.

The algorithm was tested on the SENSEVAL-2 benchmark data and shown to be on a par with the top three contenders of the SENSEVAL-2 competition. It was also shown that adding more advanced linguistic analysis to the feature extraction seems to be essential for improving the classification accuracy. We conclude that self-organized document maps have properties similar to a large-scale semantic structure that is useful for word sense disambiguation.

References

[1] T. Kohonen, S. Kaski, K. Lagus, J. Salojärvi, V. Paatero, and A. Saarela. Organization of a massive document collection. *IEEE Transactions on Neural Networks, Special Issue on Neural Networks for Data Mining and Knowledge Discovery*, 11(3):574–585, May 2000.

[2] K. Lindén and K. Lagus. Word Sense Disambiguation in Document Space. In *2002 IEEE Int. Conference on Systems, Man and Cybernetics*, Tunisia, October 6–9, 2002.

[3] K. Lindén. Word Sense Disambiguation with THESSOM. In *Workshop on Self-Organizing Maps, WSOM'03 — Intelligent Systems and Innovational Computing*, Kitakyushu, Japan, September 11–14, 2003.

[4] K. Lindén. Evaluation of Linguistic Features for Word Sense Disambiguation with Self-Organized Document Maps. Computers and the Humanities, 2004 (December).

[5] C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing.* The MIT Press, Cambridge, Massachusetts, 1999.

[6] D. Yarowsky. Unsupervised word-sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL '95)*, pages 189–196, Cambridge, MA, 1995.

10.3 Topically focusing language model

A statistical language model provides predictions for future words based on the already seen word sequence. This is important, for example, in large vocabulary continuous speech recognition (see Section 9.3) to guide the search into those phoneme sequence candidates that constitute relevant words and sentences. Especially when the vocabulary is large, say 100 000 words, the estimation of the most likely words based on the previous sequence is challenging since all possible words, let alone all word sequences, have never been seen in any data set. For example, there exist 10^{25} sequences of 5 words of a vocabulary of 100 000 words. Thus directly estimating a n:th order Markov model is generally out of the question for values of n larger than 5.

In [1] we proposed a *topically focusing language model* that is built utilizing a topical clustering of texts obtained using the WEBSOM method. The long-term dependencies [2] are taken into account by focusing the predictions of the language model according to the longer-term topical and stylistic properties of the observed speech or text.

In speech recognition suitable text data or the recognizer output can be utilized to focus the model, i.e., to select the text clusters that most closely correspond to the current discourse or topic. Next, the focused model can be applied to speech recognition or to re-rank the result hypothesis obtained by a more general model [6].

It has been previously shown that good topically organized clustering of large text collections can be achieved efficiently using the WEBSOM method (see [3]). In this project, the clustering is utilized as a basis for constructing a focusing language model. The model is constructed as follows:

Cluster a large collection of topically coherent text passages, e.g., paragraphs or short documents using the WEBSOM method. For each cluster (e.g. for each map unit), calculate a separate, small n-gram model. During speech recognition, use transcription history and the current hypothesis to select a small number of topically 'best' clusters. Combine the language models of each cluster to obtain a focused language model. This model is thus focused on the topical and stylistic peculiarities of a history of, say, 50 words. Combine further with a general language model for smoothing. The structure of the resulting combined language model is shown in Figure 10.4.

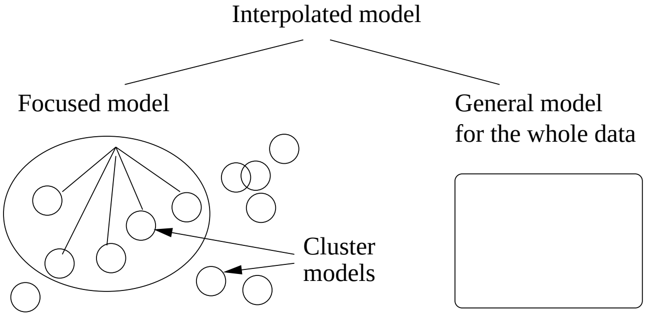


Figure 10.4: *A focusing language model obtained as an interpolation between topical cluster models and a general model.*

As the cluster-specific models and the general model we have used n-gram models of various orders. However, other types of models describing the short-term relationships between words could, in principle, be used as well. The combining operation amounts to a linear interpolation of the predicted word probabilities.

The models were evaluated using perplexity[1] on independent test data averaged over

[1] Perplexity is the inverse predictive probability for all the words in the test document.

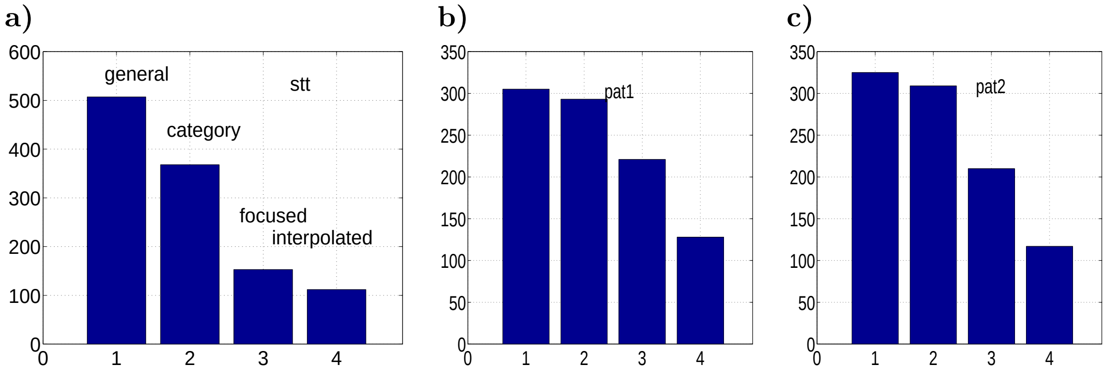


Figure 10.5: The perplexities of the different language models, **a)** for the Finnish STT news corpus, **b)** for smaller patent corpus and **c)** for larger patent corpus. The explanation of the bars in each figure, from left to right: 1. general model for the whole corpus, 2. category-specific model using prior text categories, 3. focusing model using unsupervised text clustering, and 4. the focusing model interpolated with the general model.

documents. The results for the Finnish and English text corpora in Figure 10.5 indicate that the focusing model is superior in terms of perplexity when compared to a general "monolithic" trigram model of the whole data set [4]. The focusing model is, as well, significantly better than the topic category specific models where the correct topic model was chosen based on manual class label on the data. One advantage of unsupervised topic modeling over a topic model based on fixed categories is that the unsupervised model can achieve an arbitrary granularity and a combination of several sub-topics. Finally, the lowest perplexity was obtained by a linear interpolation of word probabilities between the focusing model and the general model.

The first experiments to apply the focusing language models in Finnish large-vocabulary continuous speech recognition are reported in [5]. The results did not show clear improvements over the baseline, but by using a local LM of small but relevant text material, we see, however, that lattice rescoring can decrease the error rate. The preliminary English speech recognition tests indicate as well, that an interpolated model between a huge general LM and a small local LM performs better than the general LM alone. While there are clearly improvements to be made in language modeling, for example, to collect larger amounts of relevant text training data, maybe the most important result of the Finnish speech recognition tests is that the topical focusing works and does not slow down the whole recognition process.

More recently, our studies have been focused on conditions when the interpolation of the local LMs and the general LMs does improve the perplexity and the speech recognition results [6] and on comparisons of different combination algorithms between topic LMs and other LMs [7].

References

[1] V. Siivola, M. Kurimo, and K. Lagus. Large vocabulary statistical language modeling for continuous speech recognition in Finnish. In *Proceedings of the 7th European Conference on Speech Communication and Technology*, pages 737–730, 2001.

[2] R.M. Iyer and M. Ostendorf, "Modeling long distance dependencies in language: Topic mixtures versus dynamic cache model," *IEEE Trans. Speech and Audio Processing*, 7, 1999.

[3] M. Kurimo and K. Lagus. An efficiently focusing large vocabulary language model. In *Proceedings of the International Conference on Artificial Neural Networks (ICANN'02)*, pages 1068–1073, Madrid, Spain, 2002.

[4] K. Lagus and M. Kurimo. Language model adaptation in speech recognition using document maps. In *Proceedings of the IEEE Workshop on Neural Networks for Signal Processing*, pages 627–636, Martigny, Switzerland, 2002.

10.4 Emergence of linguistic features using independent component analysis

We have been able to show that Independent Component Analysis (ICA) [2] applied on word context data provides distinct features that reflect syntactic and semantic categories[1]. The analysis gives features or categories that are both explicit and can easily be interpreted by humans. This result can be obtained without any human supervision or tagged corpora that would have some predetermined morphological, syntactic or semantic information. We have also shown that the emergent features match well categories determined by linguists by comparing the ICA results with a tagged corpus [3].

We have also shown that the ICA can be used succesfully for studying the properties of morphemes [5]. We used a large Finnish text corpus in the analysis. As a result we obtained emergent linguistic representations for the morphemes. We have also used the ICA in language modeling. This includes creating an N-gram model for classes derived from ICA features [6].

In the following, we will show several examples of the analysis results from [1]. In considering the feature distributions, it is good to keep in mind that the sign of the features is arbitrary. As was mentioned earlier, this is because of the ambiguity of the sign: one could multiply a component by -1 without affecting the model. Also, the numbering (order) of the components is arbitrary.

Fig. 10.6 shows how the third component is strong in the case of nouns in singular form. A similar pattern was present in all the nouns with three exceptional cases with an additional strong fourth component indicated in Fig. 10.7. The reason appears to be that “psychology” and “neuroscience” share a semantic feature of being a science or a scientific discipline. A similar pattern is also present in words such as “engineering” and “biology”. This group of words provide a clear example of distributed representation where, in this case, two components are involved.

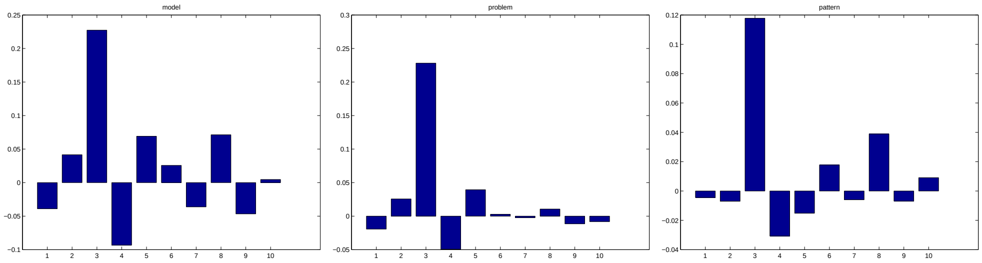


Figure 10.6: ICA features for “model”, “problem” and “pattern”. For each word, we show the values of the 10 independent components as a bar plot.

An interesting point of comparison for Fig. 10.6 is the collection of plural forms of the same nouns in Fig. 10.8. The third component is strong as with the singular nouns but now there is another strong component, the fifth.

The results include both an emergence of clear distinctive categories or features and a distributed representation. In the emergent representation, a word may thus belong to several categories simultaneously in a graded manner.

We wish that our model provides additional understanding on potential cognitive mechanisms in natural language learning and understanding [4]. Our approach attempts to show that it is possible that much of the linguistic knowledge is emergent in nature and based on specific learning mechanisms.

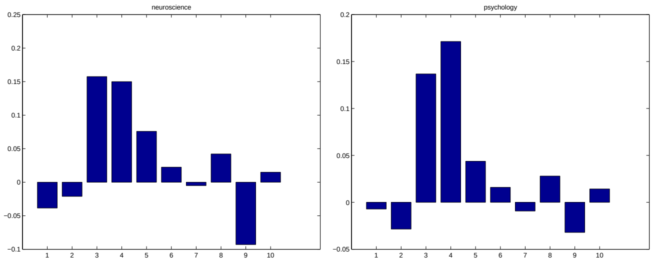


Figure 10.7: ICA features for “neuroscience” and “psychology”.

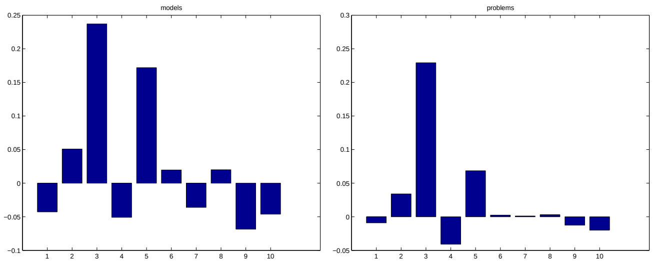


Figure 10.8: ICA features for “models” and “problems”.

References

[1] T. Honkela, and A. Hyvärinen. Linguistic Feature Extraction using Independent Component Analysis. In *Proceedings of IJCNN 2004, International Joint Conference on Neural Networks*, Budapest, Hungary, 25-29 July 2004, pp. 279-284.

[2] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis.* John Wiley & Sons, 2001.

[3] J.J. Väyrynen, T. Honkela, and A. Hyvärinen. Independent Component Analysis of Word Contexts and Comparison with Traditional Categories. In: Jarmo M. A. Tanskanen (ed.), Proceedings of the 6th Nordic Signal Processing Symposium - NORSIG 2004, Espoo, Finland, 9-11 June 2004, pp. 300-303.

[4] T. Honkela, A. Hyvärinen, and J. Väyrynen. Emergence of Linguistic Features: Independent Component Analysis of Contexts. In A. Cangelosi, G. Bugmann and R. Borisyuk (eds.), Proceedings of NCPW9, Neural Computation and Psychology Workshop, Plymouth, England, pp. 129-138, 2005.

[5] K. Lagus, M. Creutz, and S. Virpioja. Latent Linguistic Codes for Morphemes using Independent Component Analysis. In A. Cangelosi et al. (eds.), Modeling Language, Cognition and Action, Proceedings of the Ninth Neural Computation and Psychology Workshop, NCPW9, Plymouth, England, pp. 139-144.

[6] Virpioja, S. (2005) New methods for statistical natural language modeling. Master’s thesis, Department of Computer Science and Engineering, Helsinki University of Technology.

10.5 SOM-based analysis of words and sentences

Observation of language use provides indirect evidence of the representations that humans utilize. The study of conceptual and cognitive representations that underlie the use of language is important for applications such as speech recognition. By studying large amounts of data it may be possible to induce the conceptual, system-internal representations which provide a grounding for meanings of words.

The self-organizing map [4] can be applied for clustering word forms based on the words that have appeared in their immediate contexts. This has been shown originally for artificially generated sentences in [9] and later for large English text corpora, e.g., in [1]. In Finnish the rich inflectional morphology poses a challenge as the vocabularies built on inflected word forms are typically very large. This problem has successfully been tackled in [6]. In [7], the motivation and methodology of use of the self-organizing maps in conceptual analysis is considered in some detail.

In [5] we were analyzing poems, in particular, Shakespeare's sonnets. Our specific focus was to see what kind information we can find on the "semantic turn" in a sonnet. This is a topic that is related both to the structure of a poem and the meaning of the words used. Our aim was not to present an analysis model that would cover all relevant aspects but to outline one particular approach that can be later extended to cover other points of view.

In [8] we considered the creation of word category maps (cf. e.g. [9, 1]) using ICA-based word features. In earlier studies, a random encoding for each word has been used. Ideally, one could represent each word as a feature vector that would take into account its syntactic and semantic characteristics. This kind of sparse feature representation can be created automatically using independent component analysis (ICA) [3] as we have shown in [2]. In [8], we compared the word category maps both in cases where the random encoding and ICA-based encoding of words were used.

References

[1] T. Honkela. *Self-Organizing Maps in Natural Language Processing.* PhD thesis, Helsinki University of Technology, 1997, Espoo, Finland.

[2] T. Honkela, A. Hyvärinen, and J. Väyrynen. *Emergence of linguistic representation by independent component analysis.* Technical report A72, Helsinki University of Technology, Laboratory of Computer and Information Science, 2003.

[3] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis.* John Wiley & Sons, 2001.

[4] T. Kohonen. *Self-Organizing Maps.* Third, extended edition. Springer, 2001.

[5] O. Kohonen, S. Katajamäki, and T. Honkela. In Search for Volta: Statistical Analysis of Word Patterns in Shakespeare's Sonnets. In: *Proceedings of International Symposium on Adaptive Models of Knowledge, Language and Cognition (AMKLC'05).* Espoo, Finland, June 15-17, 2005. pp. 44-47.

[6] K. Lagus, A. Airola, and M. Creutz. Data analysis of conceptual similarities of Finnish verbs. In *Proceedings of the CogSci 2002, the 24th annual meeting of the Cognitive Science Society*, pages 566–571. Fairfax, Virginia, August 7–10, 2002.

[7] K. Lagus. Miten hermoverkkomallit selittävät kielen oppimista. A.M. Korpijaakko-Huuhka, S. Pekkala and H. Heimo. (eds.) *Kielen ja kognition suhde*. Puheen ja kielen tutkimuksen yhdistyksen julkaisuja 37, 2005.

[8] J.J. Väyrynen and T. Honkela. Word Category Maps based on Emergent Features Created by ICA. In: Heikki Hyötyniemi, Pekka Ala-Siuru and Jouko Seppänen (eds.), Life, Cognition and Systems Sciences, Symposium Proceedings of the 11th Finnish Artificial Intelligence Conference, Finnish Science Center Heureka Vantaa, 1-3 September 2004, pp. 173-185.

[9] H. Ritter and T. Kohonen. Self-organizing semantic maps. *Biological Cybernetics*, 1989; 61:241-254

Chapter 11

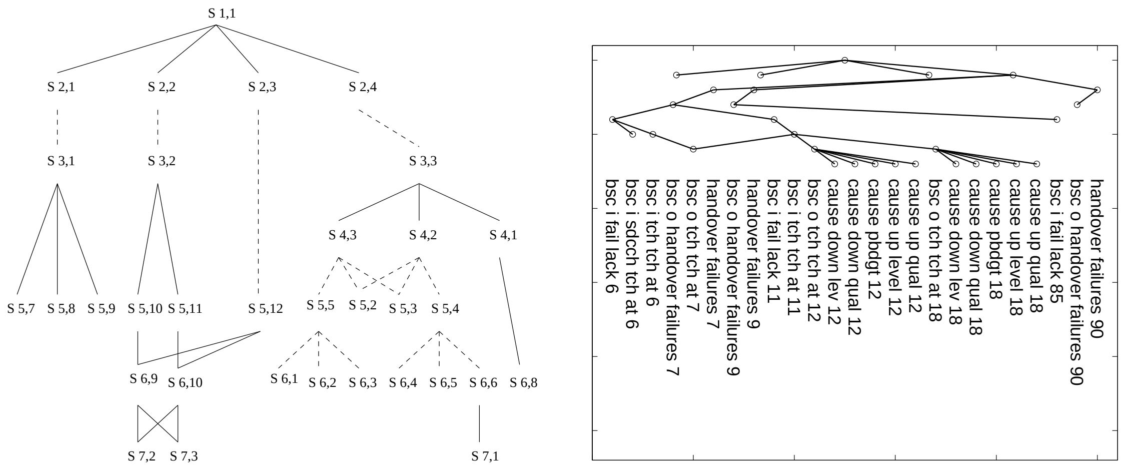
Intelligent data engineering

Olli Simula, Jaakko Hollmen, Sampsa Laine, Kimmo Raivio, Miki Sirola, Pasi Lehtimäki, Timo Similä, Mika Sulkava, Jarkko Tikka, Jukka Parviainen, Teppo Marin, Golan Lampi, Mikko Multanen, Risto Hakala, Petri Saarikko

11.1 A knowledge-based model for analysis of GSM network performance

The performance of the mobile network is measured based on thousands of counters, describing the numbers of the most important events over a measurement period (typically one hour). In order to allow more efficient performance monitoring, a set of high-level key performance indicators (KPIs) are derived from the counter data. Such indicators are traditionally used in resource management [1] and they are well suited for performance monitoring [4], but there are several drawbacks when they are used in fault diagnosis [2]. One of the main problems is that most of the widely used performance indicators describe the operation of the network at the BTS or BSC level. As a result, the performance degradations originating from interaction between several BTSs become very difficult to observe. In many cases, however, the operation of the close-by BTSs are highly dependent on each other. In this project [3], we aimed to avoid the above mentioned KPI-related problems by using a novel performance analysis approach based on counter data. Due to the significant increase in number of variables, a knowledge-based model was used to divide the analysis process into a set of small system identification problems in order to keep the overall analysis process tractable.

In order to determine the knowledge-based model, the available measurements were divided into variable sets describing the performance of the different subsystems of the GSM network. Then, a simple mathematical input-output model for each of the subsystems were proposed. In Figure 11.1(a), the subsystem hierarchy for the overall performance model is shown. The model parameters were estimated from the available data record using quadratic programming. Then, the parameter estimates were used to find the input-output variable pairs involved in the most severe performance degradations. Finally, the resulting variable pairs were visualized as a tree-shaped cause-effect chain in order to allow user friendly analysis of the network performance (see Figure 11.1(b)). The provided information can be used to enhance the current radio resource usage in the network.



(a) (b)

Figure 11.1: (a) The subsystem hierarchy. (b) The dependency tree describing the cause-effect chains of the handover failures.

References

[1] Sofoklis A. Kyriazakos and George T. Karetsos. Practical Radio Resource Management in Wireless Systems. Artech House, Inc., 2004.

[2] P. Lehtimäki and K. Raivio A SOM Based Approach for Visualization of GSM Network Performance Data. *Proceedings of the 18th Internation Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems (IEA/AIE)*, pages 588 – 598, Bari, Italy, June 22–24, 2005.

[3] P. Lehtimäki and K. Raivio A Knowledge-Based Model for Analyzing GSM Network Performance. *Proceedings of the 6th International Symposium on Intelligent Data Analysis (IDA)*, pages 205 – 215, Madrid, Spain, September 8–10, 2005.

[4] J. Laiho, K. Raivio, P. Lehtimäki, K. Hätönen, and O. Simula. Advanced analysis methods for 3G cellular networks. *IEEE Transactions on Wireless Communications*, 4(3):930–942, May 2005.

11.2 Analysis of mobile radio access network

Radio access networks produce a huge amount of data. In this project, the Self-Organizing Map has been used to analyse mobile data [1][2] from GSM and 3G networks. Earlier analysed 3G network data was generated using a radio network simulator. Now, both 3G and GSM data have been collected from real network. The goal is to develop efficient adaptive methods for monitoring the network behavior and performance. Special interest is on fault detection and on finding clusters of mobile cells or mobile cell pairs. Cells of one cluster can be configured using similar parameters.

The method utilizes the SOM algorithm several times when clustering mobile cells or cell pairs. At first, the Self-Organizing Map is used together with some clustering algorithm to cluster feature vectors of single cells or cell pairs. The clustering can be performed repeatedly to cluster mobile cells using more complex features or to zoom stepwise into some part of the data. In the latter method, at each round the most interesting subset of the data is selected for further analysis. This two phase clustering algorithm [3] begins with training a SOM with the data vectors. The codebook vectors of the SOM are clustered using K-means or some hierarchical clustering method with a validity index.

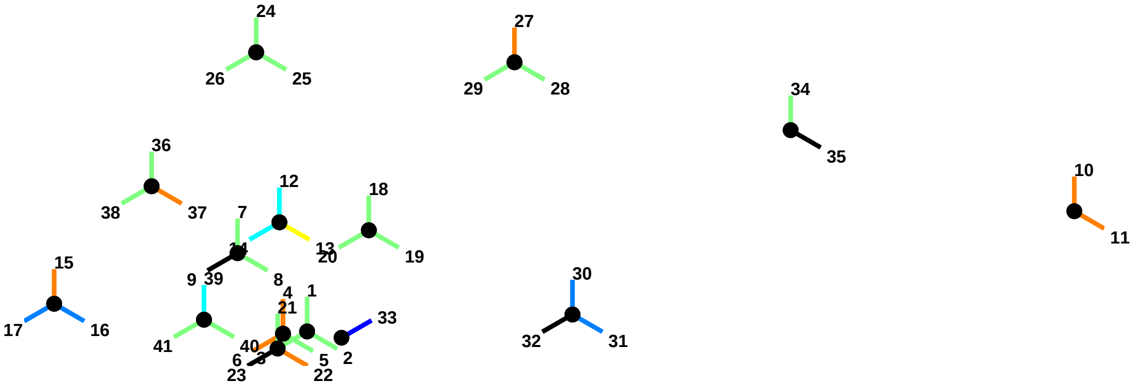


Figure 11.2: Classified GSM cells

When clustering mobile cells a histogram can be computed for each mobile cell. The histogram describes how the data from one cell fall into the data clusters. These histograms are used as profiles in cell classification. The profiles are fed into second SOM, which is clustered to find the classes of cell profiles. The classified mobile cells and their locations are presented in Figure 11.2. In this method, two level clustering procedure has been used because long term cell profiles are desired. The method gives us more reliable classification results.

References

[1] J. Laiho, K. Raivio, P. Lehtimäki, K. Hätönen, and O. Simula. Advanced analysis methods for 3G cellular networks. *IEEE Transactions on Wireless Communications*, 4(3):930–942, May 2005.

[2] P. Lehtimäki and K. Raivio. A SOM based approach for visualization of GSM network performance data. In M. Ali and F. Esposito, editors, *IEA/AIE*, volume 3533 of *Lecture Notes in Computer Science*, pages 588–598. Springer, 2005.

[3] J. Vesanto and E. Alhoniemi. Clustering of the self-organizing map. *IEEE Transactions on Neural Networks*, 11(3):586–600, May 2000.

11.3 Using visualization, variable selection and feature extraction to learn from industrial data

Co-operation between the Academia and the general society has been defined by the Finnish Parliament as a major task for Finnish Universities. One of the tactics of the Laboratory of Computer and Information Sciences to meet this challenge is the development of the HAHMO-tool depicted in Fig. 11.3. This tool is intended to facilitate the definition of the goals and realization of an applied research project. This brief description illustrates the basis of the method and the current implementation.

The HAHMO is directly based onto the CRoss Industrial Standard Platform for Data Mining (CRISP-DM), developed by Chapman et al. in [1]. The CRISP-DM is a six-phase method for the definition and realization of a statistically motivated project. The first two phases require the user to define the Business Objectives and assess the quality and quantity of data. The next two phases lead to the pre-treatment of the data and to the actual modelling. The last two phases require assessment of the business potential of the statistical results, and require a deployment plan to be drawn. The CRISP-DM has been created to support industrial data mining projects. The HAHMO is a computer interface to this analysis approach.

The HAHMO tool leads the user through the CRISP-DM -process by asking questions and allowing the user to fill in his or her answers. These answers are stored into a data base to allow later study and refinement. The HAHMO also creates a report of the data fed into the data base.

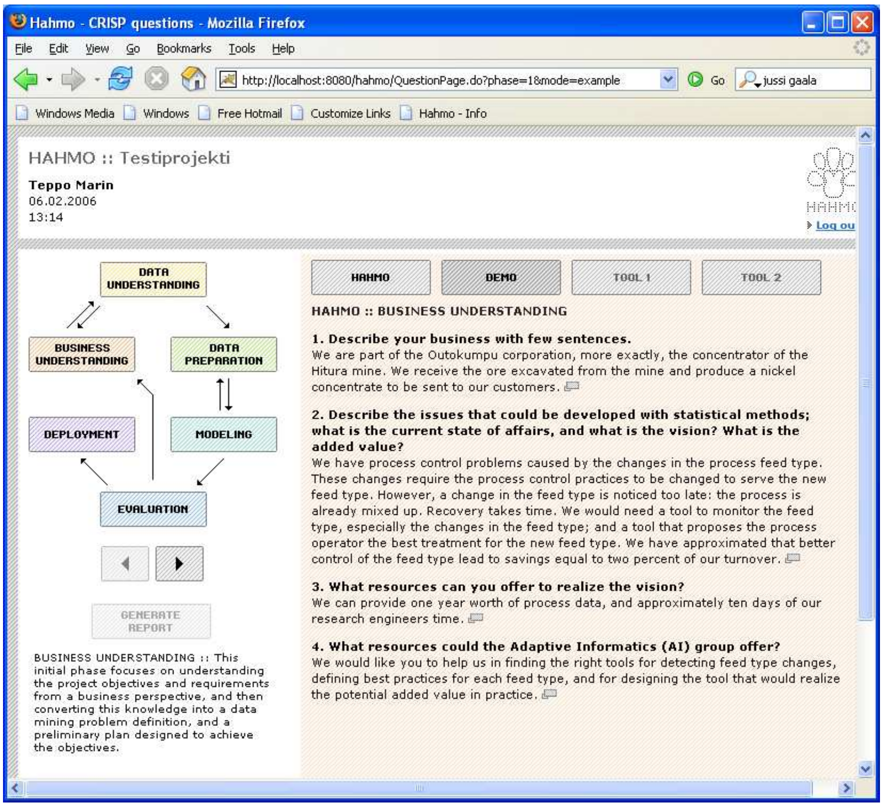
The HAHMO has been, so far, piloted once. In Autumn 2005 we invited Isto Halinen from Efeko to pilot the HAHMO-tool. We held a one hour meeting in which we used the HAHMO to define their needs and our capability of serving those needs. The benefits of the HAHMO were as follows:

1. the meeting proceeded in a well organized manner, and
2. all relevant aspects of the task were traversed.

Mr. Halinen was pleased with the results and intended to forward the created plan in Kunnallisliitto, an organization to which they are related.

References

[1] P. Chapman, J. Clinton, T. Khabaza, T. Reinartz, and R. Wirth CRISP-DM 1.0 step-by-step data mining guide. *Technical report, CRISM-DM consortium, http://www.crisp-dm.org*, 2000.


Hahmo - CRISP questions - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://localhost:8080/hahmo/QuestionPage.do?phase=1&mode=example
Go
jussi gaala
Windows Media Windows Free Hotmail Customize Links Hahmo - Info
HAHMO
Log ou
DATA UNDERSTANDING
BUSINESS UNDERSTANDING
DATA PREPARATION
DEPLOYMENT
MODELING
EVALUATION
GENERATE REPORT
HAHMO
DEMO
TOOL 1
TOOL 2

HAHMO :: Testiprojekti

Teppo Marin
06.02.2006
13:14

HAHMO :: BUSINESS UNDERSTANDING

1. Describe your business with few sentences.
We are part of the Outokumpu corporation, more exactly, the concentrator of the Hitura mine. We receive the ore excavated from the mine and produce a nickel concentrate to be sent to our customers.

2. Describe the issues that could be developed with statistical methods; what is the current state of affairs, and what is the vision? What is the added value?
We have process control problems caused by the changes in the process feed type. These changes require the process control practices to be changed to serve the new feed type. However, a change in the feed type is noticed too late: the process is already mixed up. Recovery takes time. We would need a tool to monitor the feed type, especially the changes in the feed type; and a tool that proposes the process operator the best treatment for the new feed type. We have approximated that better control of the feed type lead to savings equal to two percent of our turnover.

3. What resources can you offer to realize the vision?
We can provide one year worth of process data, and approximately ten days of our research engineers time.

4. What resources could the Adaptive Informatics (AI) group offer?
We would like you to help us in finding the right tools for detecting feed type changes, defining best practices for each feed type, and for designing the tool that would realize the potential added value in practice.

BUSINESS UNDERSTANDING :: This initial phase focuses on understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition, and a preliminary plan designed to achieve the objectives.

Figure 11.3: A screen shot from the HAHMO-tool.

11.4 SOM in decision support

The usability of Self-Organizing Map (SOM) method in computerized decision support systems is studied [1]. A prototype (DERSI) that combines neural methods and knowledge-based methodologies is being built (Fig. 11.4). DERSI is intended to be used first in fault diagnosis, but also other application areas are possible. Decision making problems in fault detection and identification are met for instance in the control rooms of power plants.

One goal of the study is to develop the control room tool with co-operation of end-users and other specialists. In addition the intention is to make data analysis with real data and simulated data of nuclear power plants, and analyze various failures. Some co-operation with the TVO Olkiluoto and Finnish Radiation Centre (STUK) have been initialized. One set of TVO simulator data is already being analyzed. One important research issue is to find out in what kind of failures the analysis can really help to find out the reason for the fault early enough, and concentrate on such scenarios in more detailed analysis. In initial studies e.g. leaks were found out promising in this respect.

DERSI is a general framework that can be used also in other purposes and application areas. For instance, paper industry could be a potential process where the same framework could be utilized.

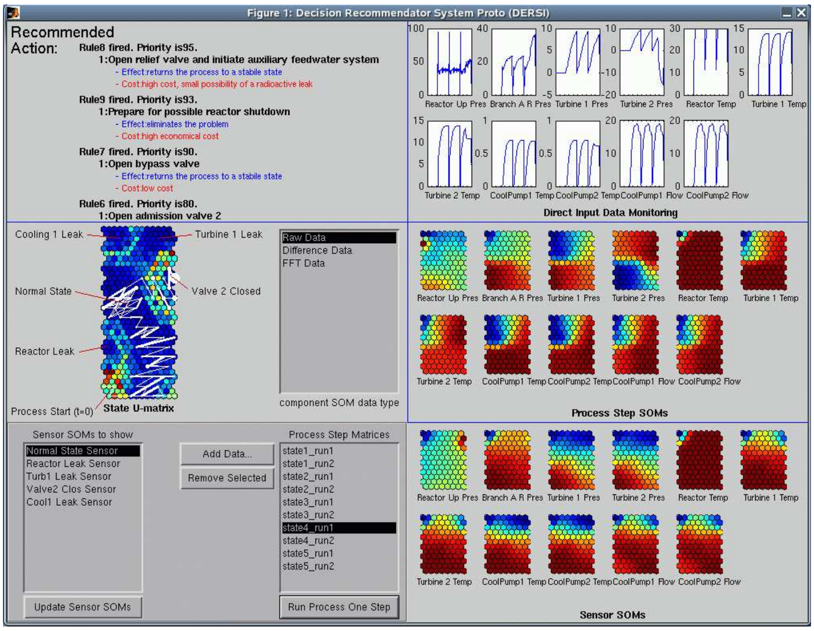


Figure 1: Decision Recommendator System Proto (DERSI)

Recommended Action:

Rule8 fired. Priority is95.
1:Open relief valve and initiate auxiliary feedwater system
- Effect:returns the process to a stabile state
- Cost:high cost, small possibility of a radioactive leak

Rule9 fired. Priority is93.
1:Prepare for possible reactor shutdown
- Effect:eliminates the problem
- Cost:high economical cost

Rule7 fired. Priority is90.
1:Open bypass valve
- Effect:returns the process to a stabile state
- Cost:low cost

Rule6 fired. Priority is80.
1:Open admission valve 2

Figure 11.4: DERSI Man-Machine Interface (MMI) including different decision support visualizations.

References

[1] M. Sirola, G. Lampi, J. Parviainen. SOM based decision support in failure management. *International Journal of Computing*, 4(3): 124–130, 2005.

11.5 Interpreting dependencies in data using the Self-Organizing Map

The Self-Organizing Map (SOM) is a widely used method for the visualization of multivariate data. This work presents various applications of the SOM.

Model interpretation. Traditional regression means predicting the mean value of the output for a given value of the input. Instead of the mean value, quantile regression tries to predict the median or any other quantile of the output given the input. In [1] nonlinear quantile regression models and their interpretation is considered. Model interpretation is important if the purpose is not only the prediction itself, but the aim is to understand dependencies between the inputs and outputs. A novel model visualization technique is proposed based on the SOM.

Variable selection. Relations between two distinct sets of multivariate data are studied in [2, 3]. The SOM is used to visualize the dependent set of data. Then the proposed method finds variables of the independent set that are related to the visualization. To illustrate the method, we applied it to the task of finding the common properties of various car models that explain their safety and economic aspects.

Manifold learning. A modification of the SOM algorithm for visualizing a special type of data is presented [4]. It is known that the SOM has problems with data, which form a nonlinear manifold (for instance a heavily twisted sheet) in a high-dimensional space. On the other hand, some of the more recently presented projection methods can handle these cases. The proposed technique is a hybrid of the traditional SOM and one of these recent methods, see Fig. 11.5. Although the technique may seem to have a narrow area of applicability, there are problems, for instance in the area of image analysis, where intrinsically low dimensional and nonlinear data structures lie in a very high-dimensional pixel space.

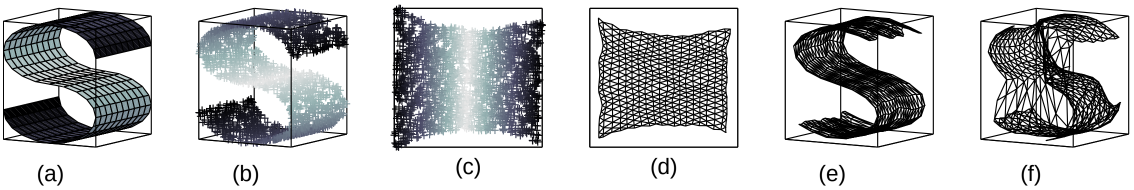


Figure 11.5: The problem of manifold learning for three-dimensional data (b) sampled from an intrinsically two-dimensional S-curve manifold (a). The Locally Linear Embedding algorithm [5] discovers the projection onto the internal coordinates on the manifold (c). The map trained by the algorithm proposed in [4] learns similarities in the internal coordinates (d) and forms a successful representation of the manifold in the observation coordinates (e). The basic SOM algorithm fails in this task (f).

References

[1] T. Similä. Self-organizing map visualizing conditional quantile functions with multidimensional covariates. *Computational Statistics & Data Analysis*, 50(8):2097–2110, 2006.

[2] S. Laine and T. Similä. Using SOM-based data binning to support supervised variable selection. *Proceedings of the 11th International Conference on Neural Information Processing (ICONIP)*, pages 172–180, Calcutta, India. November 22–25, 2004.

[3] T. Similä and S. Laine. Visual approach to supervised variable selection by self-organizing map. *International Journal of Neural Systems*, 15(1–2):101–110, 2005.

[4] T. Similä. Self-organizing map learning nonlinearly embedded manifolds. *Information Visualization*, 4(1):22–31, 2005.

[5] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

11.6 Analysis of forest nutrition data

Forests are complex ecosystems. Gaining an insight into the condition of forests and the assessment of the future development of forests under the present and predicted environmental scenarios requires large data sets from long-term monitoring programmes. In this project the development of forests in Finland has been studied using data from the International Cooperative Programme on the Assessment and Monitoring of Air Pollution Effects on Forests (ICP Forests).

Plant nutrients play an integral role in the physiological and biochemical processes of forest ecosystems. We have analyzed the development of foliar nutrient concentrations in coniferous trees in Finland using clustering of the Self-Organizing Map [1, 2]. Based on these results further analysis of the effect of nitrogen and sulfur deposition on the mineral composition of foliage was conducted [3]. It was concluded that evidence for deposition-induced changes in needles has clearly decreased during the nineties.

Various environmental factors and past development affect the growth and nutritional composition of tree needles as they are aging. Different regression models were compared to find out how these effects [4] could be modelled effectively and accurately during the second year of the needles. We found that sparse regression models are well suited for this kind of analysis. They are better for the task than ordinary least squares single and multiple regression models, because they are both easy to interpret and accurate at the same time. Average R^2 values of the different regression models for different element concentrations and mass of the needles (NM) are presented in Figure 11.6.

Good quality of analytical measurements techniques is important to ensure the reliability of analyses in environmental sciences. We combined foliar nutrition data from Finland and results of multiple measurement quality tests from different sources in order to study the effect of measurement quality on conclusions based on foliar nutrient analysis [5]. In particular, we studied the use of weighted linear regression models in detecting trends in foliar time series data and showed that good precision of the measurement techniques may decrease the time needed to detect statistically significant trends in environmental time series by several years.

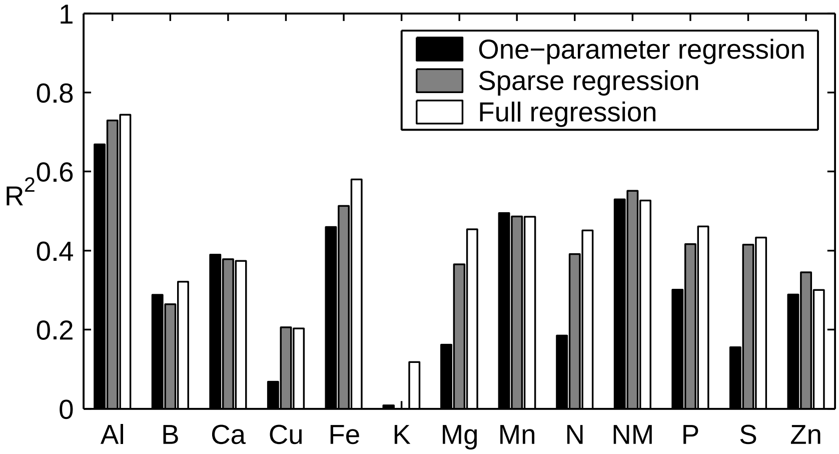


Figure 11.6: *Average R^2-values of one-parameter regression, sparse regression and full regression models for foliar measurements from pine needles obtained using cross-validation. Results are for validation sets.*

References

[1] Juha Vesanto and Mika Sulkava. Distance matrix based clustering of the self-organizing map. In José R. Dorronsoro, editor, *Artificial Neural Networks - ICANN 2002: International Conference, Proceedings*, volume 2415 of *Lecture Notes in Computer Science*, pages 951–956, Madrid, Spain, August 2002. Springer-Verlag.

[2] Sebastiaan Luyssaert, Mika Sulkava, Hannu Raitio, and Jaakko Hollmén. Evaluation of forest nutrition based on large-scale foliar surveys: are nutrition profiles the way of the future? *Journal of Environmental Monitoring*, 6(2):160–167, February 2004.

[3] Sebastiaan Luyssaert, Mika Sulkava, Hannu Raitio, and Jaakko Hollmén. Are N and S deposition altering the chemical composition of Norway spruce and Scots pine needles in Finland? *Environmental Pollution*, 138(1):5–17, November 2005.

[4] Mika Sulkava, Jarkko Tikka, and Jaakko Hollmén. Sparse regression for analyzing the development of foliar nutrient concentrations in coniferous trees. In Sašo Džeroski, Bernard Ženko, and Marko Debeljak, editors, *Proceedings of the Fourth International Workshop on Environmental Applications of Machine Learning (EAML 2004)*, pages 57–58, Bled, Slovenia, September/October 2004.

[5] Mika Sulkava, Pasi Rautio, and Jaakko Hollmén. Combining measurement quality into monitoring trends in foliar nutrient concentrations. In Włodzisław Duch, Janusz Kacprzyk, Erkki Oja, and Sławomir Zadrożny, editors, *Artificial Neural Networks: Formal Models and Their Applications - ICANN 2005: 15th International Conference, Proceedings, Part II*, volume 3697 of *Lecture Notes in Computer Science*, pages 761–767, Warsaw, Poland, September 2005. Springer-Verlag.

11.7 Parsimonious signal representations in data analysis

Data mining is the analysis of often large observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the owner [1]. While utility is a natural starting point for any analysis, understandability often remains a secondary goal. In this research, improved understandability of data-analytic models is sought by investigating sparse signal representations that are learned from data.

In order to learn about the relationships between variables, a method for learning parsimonious dependency structures has been presented in [4]. The class of dependency structures is the set of linear regression models. Each variable is taken as the output variable in turn and a sparse regression algorithm is used to select a parsimonious set of inputs from the rest of variables. A bootstrap-based procedure has been developed in order estimate robust dependency structures. A linear dependency forest of System data is shown in Figure 11.7. The System data consist of measurements from nine variables of a single computer which is connected to a network [7]. In Fig. 11.7, `idle`, `ipkts`, and `blks/s` are the output variables and the rest are the input variables. The markers `+` and `-` indicate positive and negative effect of the input variable on the output variable, respectively.

In a time series context, parsimonious modeling techniques can be used in estimating a sparse set of autoregressive variables for time series prediction [5]. We present an algorithm in the spirit of backward selection, which removes variables sequentially from the prediction models based on the significance of the individual regressors using bootstrap-based confidence intervals for the prediction error.

In ecology, a needle aging prediction problem has been casted into the framework of parsimonious modeling. In [2, 3], we show how linear sparse regression models can be used to represent the relations between different foliar nutrient concentration measurements of coniferous trees in consecutive years. In the experiments, the models proved to be capable of providing relatively good and reliable predictions of the development of foliage with a considerably small number of regressors. Two methods for estimating sparse models were compared to more conventional linear regression models. Differences in the prediction accuracies between the sparse and full models were minor, but the sparse models were found to highlight important dependencies between the nutrient measurements better than the other regression models. The use of sparse models is, therefore, advantageous in the analysis and interpretation of the development of foliar nutrient concentrations.

The problem of estimating sparse regression models in a case of multi-dimensional input and output variables has been investigated in [6]. A forward-selection algorithm that extends the Least Angle Regression algorithm (LARS) is presented. The proposed method is also applied to the task of selecting relevant pixels from images in multidimensional scaling of handwritten digits.

The research work on parsimonious modeling will be continued in various applications in bioinformatics, ecology and general data mining problems. The projects involving domain expertise will be especially useful to see the how well the results can be interpreted.

References

[1] David Hand, Heikki Mannila, and Padhraic Smyth. *Principles of Data Mining.* Adaptive Computation and Machine Learning Series. MIT Press, 2001.

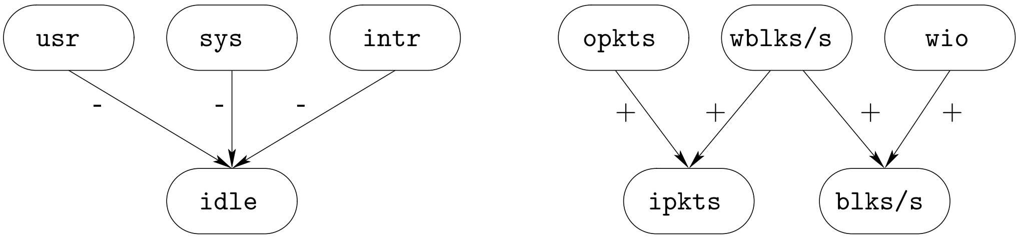


Figure 11.7: The linear dependency forest obtained from the System data. The arrows point from the input variables to the output variables.

[2] Mika Sulkava, Jarkko Tikka, and Jaakko Hollmén. Sparse regression for analyzing the development of foliar nutrient concentrations in coniferous trees. In Sašo Džeroski, Bernard Ženko, and Marko Debeljak, editors, *Proceedings of the Fourth International Workshop on Environmental Applications of Machine Learning (EAML 2004)*, pages 57–58, 2004.

[3] Mika Sulkava, Jarkko Tikka, and Jaakko Hollmén. Sparse regression for analyzing the development of foliar nutrient concentrations in coniferous trees. *Ecological Modeling*, 191(1):118–130, 2006.

[4] Jarkko Tikka and Jaakko Hollmén. Learning linear dependency trees from multivariate time-series data. In *Proceedings of the Workshop on Temporal Data Mining: Algorithms, Theory and Applications (in conjunction with The Fourth IEEE International Conference on Data Mining)*, Brighton, U.K., 2004.

[5] Jarkko Tikka, Jaakko Hollmén, and Amaury Lendasse. Input Selection for Long-Term Prediction of Time Series. In Joan Cabestany, Alberto Prieto, and Francisco Sandoval, editors, *Proceedings of the 8th International Work-Conference on Artificial Neural Networks (IWANN 2005)*, volume 3512 of *Lecture Notes in Computer Science*, pages 1002–1009. Springer-Verlag, June 2005. Vilanova i la Geltú, Barcelona, Spain.

[6] Timo Similä and Jarkko Tikka. Multiresponse Sparse Regression with Application to Multidimensional Scaling. In Wlodzislaw Duch, Janusz Kacprzyk, Erkki Oja, and Slawomir Zadrozny, editors, *Proceedings of the 15th International Conference on Artificial Neural Networks (ICANN)*, volume 3967 part II of *LNCS, Springer-Verlag*, pages 97–102, Warsaw, Poland, September 2005.

[7] Juha Vesanto and Jaakko Hollmén. An Automated Report Generation Tool for the Data Understanding Phase. In Ajith Abraham, Lakhmi Jain, and Berend J. van der Zwaag, editors, *Innovations in Intelligent Systems: Design, Management and Applications*, Springer (Physica) Verlag, Studies in Fuzziness and Soft Computing, 2003.

Chapter 12

Time series prediction

Amaury Lendasse, Yongnan Ji, Nima Reyhani, Jin Hao, Antti Sorjamaa

12.1 Introduction

Amaury Lendasse

What is Time series prediction? Time series prediction (TSP) is a challenge in many fields. In finance, experts forecast stock exchange courses or stock market indices; data processing specialists forecast the flow of information on their networks; producers of electricity forecast the load of the following day. The common point to their problems is the following: how can one analyse and use the past to predict the future? Many techniques exist: linear methods such as ARX, ARMA, etc., and nonlinear ones such as artificial neural networks. In general, these methods try to build a model of the process. The model is then used on the last values of the series to predict the future values. The common difficulty to all the methods is the determination of sufficient and necessary information for an accurate prediction.

A new challenge in the field of time series prediction is the Long-Term Prediction: several steps ahead have to be predicted. Long-Term Prediction has to face growing uncertainties arising from various sources, for instance, accumulation of errors and the lack of information.

Our contributions in TSP research. The TSP group is a new research group. It has been created in 2004. A notable achievement has been the organization of a time series prediction challenge and the creation of a new benchmark in the field "The Cats Benchmark" (`http://www.cis.hut.fi/ lendasse/competition/competition.html`).

In the reporting period 2004 - 2004, TSP research has been established as a new project in the laboratory. Nevertheless, TSP research has already been extended to a new direction: "Chemometry".

This Chapter starts by introducing some theoretical advances undertaken during the reporting period, including the presentation and the results of the CATS Benchmark. Also the problem of input selection for TSP is reported. The applications range includes Chemometry.

12.2 The CATS benchmark

Amaury Lendasse

In the CATS competition [2], the goal was the prediction of 100 missing values of the time series; they are grouped in 5 sets of 20 successive values. The prediction methods have then to be applied several times, allowing a better comparison of the performances. Twenty-four papers and predictions were submitted to the competition (organized during IJCNN'04). Seventeen papers were accepted according to the quality of the prediction and the quality of the paper itself. Seven papers have been accepted for oral presentation and ten for poster presentation.

This series is represented in Fig. 12.1. This artificial time series is given with 5,000 data, among which 100 are missing. The missing values are divided in 5 blocks:

- elements 981 to 1,000;
- elements 1,981 to 2,000;
- elements 2,981 to 3,000;
- elements 3,981 to 4,000;
- elements 4,981 to 5,000.

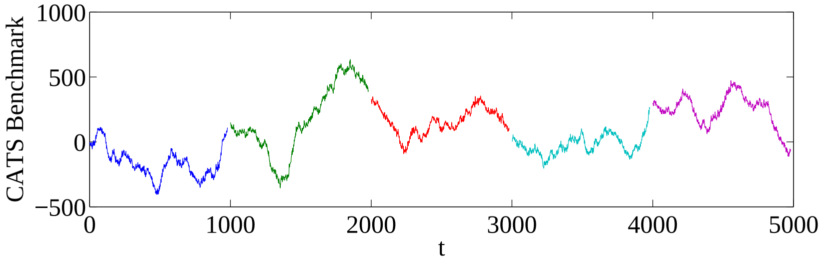


Figure 12.1: The CATS Benchmark.

The Mean Square Error is computed on the 100 missing values. The 24 methods that were submitted to the competition are very different and give very dissimilar results. The Error is in a range between 408 and 1714. It is important to notice that some methods are very good for the prediction of the eighty first values but very bad for the last 20 ones. The method that has been used by the winner of the competition is divided in two parts: the first sub-method provides the short-term prediction and the second sub-method provides the long-term one. Both sub-methods are linear, but according to the author better results could be obtained if the first sub-method was nonlinear. According to this author, the key of a good prediction is this division between two subproblems.

12.3 Methodology for long-term prediction of time series

Amaury Lendasse, Yongnan Ji, Nima Reyhani, Jin Hao, Antti Sorjamaa

The time series prediction problem is the prediction of future values based on the previous values and the current value of the time series (see Equation 12.1).

$$\hat{y}_{t+1} = f_1(y_t, y_{t-1}, ..., y_{t-M+1}). \tag{12.1}$$

The previous values and the current value of the time series are used as inputs for the prediction model. One-step ahead prediction is needed in general and is referred as Short-Term Prediction. But when multi-step ahead predictions are needed, it is called Long-Term Prediction problem.

Unlike the Short-Term time series prediction, the Long-Term Prediction is typically faced with growing uncertainties arising from various sources. For instance, the accumulation of errors and the lack of information make the prediction more difficult. In Long-Term Prediction, performing multiple steps ahead prediction, there are several alternatives to build models. Two variants of prediction strategies are studied and compared: the Direct and the Recursive Prediction Strategies [5].

12.4 Input selection strategies

Amaury Lendasse, Yongnan Ji, Nima Reyhani, Jin Hao, Antti Sorjamaa

Input selection is an essential pre-processing stage to guarantee high accuracy, efficiency and scalability in problems such as machine learning, especially when the number of observations is relatively small compared to the number of inputs. It has been the subject in many application domains like pattern recognition, process identification, time series modeling and econometrics. Problems that occur due to poor selection of input variables are:

- If the input dimensionality is too large, the 'curse of dimensionality' problem may happen. Moreover, the computational complexity and memory requirements of the learning model increase. Additional unrelated inputs lead to poor models (lack of generalization).
- Understanding complex models (too many inputs) is more difficult than simple models (less inputs), which can provide comparable good performances.

In the TSP group, two input selection methods based on different criteria have been studied: Mutual Information and Nonparametric Noise Estimator [3, 4, 6, 7, 8].

12.5 Chemometry

Many analytical problems related to spectrometry require predicting a quantitative variable through a set of measured spectral data. For example, one can try to predict a chemical component concentration in a product through its measured infrared spectrum. In recent years, the importance of such problems in various fields including the pharmaceutical, food and textile industries have grown dramatically. The chemical analysis by spectrophotometry rests on the fast acquisition of a great number of spectral data (several hundred, even several thousands).

In spectrometric problems, one is often faced with databases having more variables (spectra components) than samples; and almost all models use at least as many parameters as the number of input variables. These two problems, colinearity and risk of overfitting, already exist in linear models. However, their effect may be even more dramatic when nonlinear models are used (there are usually more parameters than in linear models, and the risk of overfitting is higher). In such high-dimensional problems, it is thus necessary to use a smaller set of variables than the initial one. We have proposed methods to select spectral variables by using two concept from information theory: the measure of mutual information [1] and the nonparametric noise estimation.

References

[1] Fabrice Rossi, Amaury Lendasse, Damien François, Vincent Wertz, and Michel Verleysen. Mutual information for the selection of relevant variables in spectrometric nonlinear modelling, Chemometrics and Intelligent Laboratory Systems, Volume 80, Issue 2, 15 February 2006, pages 215-226.

[2] Amaury Lendasse, Erkki Oja, Olli Simula, and Michel Verleysen, Time Series Prediction Competition: The CATS Benchmark, *Proceedings of IJCNN 2004*, International Joint Conference on Neural Networks, Budapest (Hungary), vol. II, 25-29 July 2004, pages 1615–1620.

[3] Antti Sorjamaa, Jin Hao, and Amaury Lendasse. Mutual Information and k-Nearest Neighbors approximator for Time Series Predictions. In *Lecture Notes in Computer Science, Proceedings of ICANN 2005*, Publisher: Springer-Verlag GmbH, volume 3697, pages 553–558, Warsaw, Poland, September 2005.

[4] Amaury Lendasse, Yongnan Ji, Nima Reyhani, and Michel Verleysen. LS-SVM Hyperparameter Selection with a Nonparametric Noise Estimator. In *Lecture Notes in Computer Science, Proceedings of ICANN 2005*, Publisher: Springer-Verlag GmbH, volume 3697, pages 625–630, Warsaw, Poland, September 2005.

[5] Yongnan Ji, Jin Hao, Nima Reyhani, and Amaury Lendasse. Direct and Recursive Prediction of Time Series Using Mutual Information Selection. In *Lecture Notes in Computer Science, Proceedings of IWANN 2005*, Publisher: Springer-Verlag GmbH, volume 3512, pages 1010–1017, July 2005.

[6] Antti Sorjamaa, Nima Reyhani, and Amaury Lendasse. Input and Structure Selection for k-NN Approximator. In *Lecture Notes in Computer Science, Proceedings of IWANN 2005*, Publisher: Springer-Verlag GmbH, volume 3512, pages 985–991, July 2005.

[7] Antti Sorjamaa, Amaury Lendasse, and Michel Verleysen. Pruned Lazy Learning Models for Time Series Prediction. In *Proceedings of ESANN 2005*, pages 509–514, Bruges, Belgium, April 2005.

[8] Nima Reyhani, Jin Hao, Yongnan Ji, and Amaury Lendasse. Mutual Information and Gamma Test for Input Selection. In *Proceedings of ESANN 2005*, pages 503–508, Bruges, Belgium, April 2005.

Chapter 13

Proactive information retrieval

Samuel Kaski, Jarkko Salojärvi, Eerika Savia, and Kai Puolamäki

13.1 Introduction

Successful proactivity, i.e. anticipation, in varying contexts requires generalization from past experience. Generalization, on its part, requires suitable powerful (stochastic) models and a collection of data about relevant past history to learn the models.

The goal of the PRIMA (Proactive Information Retrieval by Adaptive Models of Users' Attention and Interests) project is to build statistical machine learning models that learn from the actions of people to model their intentions and actions. The models are used for disambiguating the users' vague commands and anticipating their actions.

Our application area is information retrieval, where we investigate to what extent the laborious explicit relevance feedback can be complemented or even replaced by implicit feedback derived from patterns of eye fixations and movements that exhibit both voluntary and involuntary signs of users intentions. Inference is supported by models of document collections and interest patterns of users.

PRIMA is a consortium with Complex Systems Computation Group, Helsinki Institute for Information Technology (Prof. Petri Myllymäki), and Center for Knowledge and Innovation Research (CKIR), Helsinki School of Economics (Doc. Ilpo Kojo). The project lasted for 2003 – 2005.

13.2 Implicit relevance feedback from eye movements

A promising new source of implicit feedback is eye movements measured during reading. During complex tasks such as reading, attention approximately lies on the location of the reader's gaze. Therefore the eye movements should contain information on the reader's interests. Deriving interest from a reading pattern is difficult however, since the signal is complex and very noisy, and interestingness or relevance is higly subjective and thus hard to define.

In a first feasibility study we constructed a controlled experimental setting in which it is known which documents are relevant [1]. The user was instructed to find an answer to a specific question, and was then shown a set of ten document titles (Fig. 13.1), of which four were known to be relevant (i.e. thay handled the same topic as the question), and one was the correct answer. Eye movements were recorded while the users were seeking the answer. Based on data gathered from eleven test subjects we then learned models that predict relevance using the measured eye movement patterns. The study was the first evidence (with statistical significance) that inferring relevance is possible [2].

The data was used in a Pascal NoE challenge "Inferring Relevance from Eye Movements", during March 2005 – October 2005. The challenge was organized in the form of a competition where the winner was the team that could most accurately predict relevance for a left-out test data set. The results of the challenge were reported in a workshop at Neural Information Processing Systems (NIPS) in December 2005.

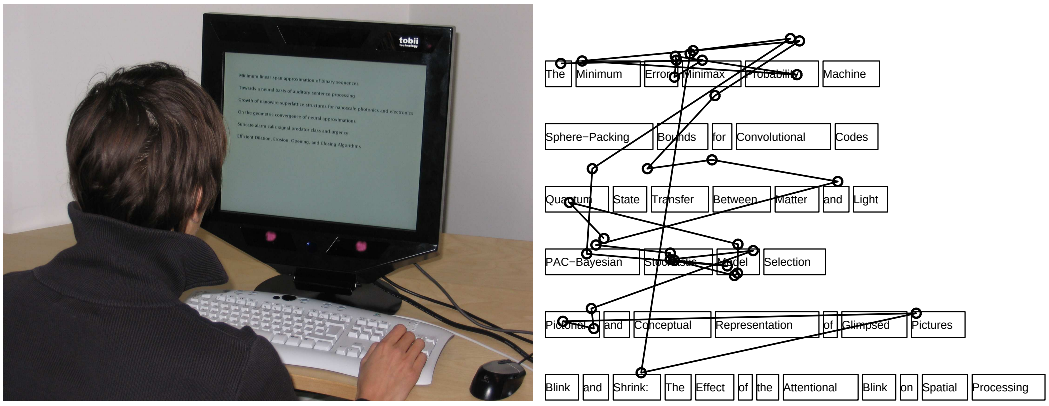

tobii
The Minimum Error Minimax Probability Machine
Sphere−Packing Bounds for Convolutional Codes
Quantum State Transfer Between Matter and Light
PAC−Bayesian Stochastic Model Selection
Pictorial and Conceptual Representation of Glimpsed Pictures
Blink and Shrink: The Effect of the Attentional Blink on Spatial Processing

Figure 13.1: The experimental setup. Left: The eye movements of the user are being tracked with Tobii 1750 eye tracker. The tracker consists of an infra-red light source and a camera integrated into the monitor frame. Right: An example of an eye movement pattern during reading, plotted on the assignment. Lines connect successive fixations, denoted by circles (Matlab reconstruction). Each line contains one document title, and some of the titles are known to be relevant.

References

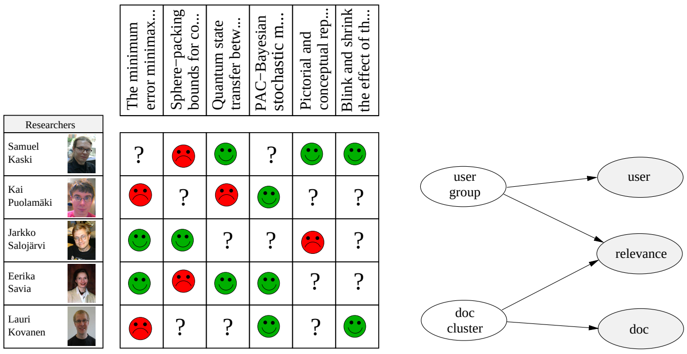
[1] Jarkko Salojärvi, Ilpo Kojo, Jaana Simola, and Samuel Kaski. Can relevance be inferred from eye movements in information retrieval? In *Proceedings of WSOM'03, Workshop on Self-Organizing Maps*, pages 261–266. Kitakyushu, Japan, 2003.

[2] Jarkko Salojärvi, Kai Puolamäki, and Samuel Kaski. Implicit relevance feedback from eye movements. In *Artificial Neural Networks: Biological Inspirations – ICANN 2005*, Lecture Notes in Computer Science 3696, pages 513–518, Berlin, Germany, 2005.

13.3 Collaborative Filtering: Inferring user interests from other available sources

Traditionally, user preferences have been predicted using so-called collaborative filtering methods, where the predictions are based on the opinions of similar-minded users. Collaborative filtering is needed when the task is to make personalized predictions but there is not yet sufficient amount of data about the user's personal interests [2]. Then the only possibility is to generalize over users, for instance by grouping them into like-minded user groups. The early methods were memory-based; predictions were made by identifying a set of similar users, and using their preferences fetched from memory. Model-based approaches are justified by the exponentially increasing time and memory requirements of the memory-based techniques. Recent work includes probabilistic and information-theoretic models. An interesting family of models are the latent component models, which have been successfully used in collaborative filtering. In these models, each user is assumed to belong to one or many latent user groups that explain her preferences. We went one step further and introduced a similar latent structure for the documents as well.

As a collaborative filtering system has to rely on the past experiences of the users, it will have problems when assessing new documents not seen yet by most of the users. To tackle this problem we have introduced a novel latent grouping model for predicting the relevance of a new document to a user [1]. The model assumes a latent group structure for both users and documents. See Figure 13.2.



Researchers	The minimum error minimax...	Sphere−packing bounds for co...	Quantum state transfer betw...	PAC−Bayesian stochastic m...	Pictorial and conceptual rep...	Blink and shrink the effect of th...
Samuel Kaski	?	☹	🙂	?	🙂	🙂
Kai Puolamäki	☹	?	☹	🙂	?	?
Jarkko Salojärvi	🙂	🙂	?	?	☹	?
Eerika Savia	🙂	☹	🙂	🙂	?	?
Lauri Kovanen	☹	?	?	🙂	?	🙂

Figure 13.2: Left: An example of gathered data for collaborative filtering. Right: The model generalizes both over users and documents.

We compared the model against a state-of-the-art method, the User Rating Profile model, where only users have a latent group structure. We estimate both models by Gibbs sampling. The new method predicts relevance more accurately for new documents that have few known ratings. The reason is that generalization over documents then becomes necessary and hence the two-way grouping is profitable.

References

[1] E. Savia, K. Puolamäki, J. Sinkkonen, and S. Kaski. Two-way latent grouping model for user preference prediction. In F. Bacchus and T. Jaakkola, editors, *Uncertainty in Artificial Intelligence 21*, pages 518–525. AUAI Press, Corvallis, Oregon, 2005.

[2] Eerika Savia, Samuel Kaski, Ville Tuulos, and Petri Myllymäki. On text-based estimation of document relevance. In *Proc. IJCNN'04*, pages 3275–3280, 2004.

13.4 Application: Proactive information retrieval

We have studied a new task, proactive information retrieval by combining implicit relevance feedback and collaborative filtering [1]. We constructed a controlled experimental setting, in which the users tried to find interesting scientific articles by browsing their titles. The experimental setup was designed to resemble closely a real-world information retrieval scenario, where the user browses the output of, e.g., a web search engine in an attempt to find interesting documents. The test subjects rated the relevance of titles of scientific articles and eye movements were measured from a subset of the test subjects.

Implicit feedback was inferred from eye movement signals, with discriminative hidden Markov models estimated from existing data in which explicit relevance feedback was available. It produced a reasonable, though rather noisy, prediction of document relevance based on eye movement measurements.

We complemented the eye movement based prediction with a probabilistic collaborative filtering model that produced a quite robust document relevance prediction. The model was computed using Markov Chain Monte Carlo techniques.

In our scenario a database of user preferences was combined with the measured implicit relevance feedback, resulting in more accurate relevance predictions. We introduced a probabilistic mixture model [1, 2] that can be used to combine the predictions. The mixture model clearly outperformed a simple linear method and was found necessary for making use of several information sources, the quality of which varied. The best prediction accuracy still leaves room for improvement but shows that proactive information retrieval and combination of many sources of relevance feedback is feasible.

Our work provides the next step towards proactive information retrieval systems. A future extension is to supplement or replace the eye movements by other sources of implicit feedback.

References

[1] Kai Puolamäki, Jarkko Salojärvi, Eerika Savia, Jaana Simola, and Samuel Kaski. Combining eye movements and collaborative filtering for proactive information retrieval. In G. Marchionini, A. Moffat, J. Tait, R. Baeza-Yates, and N. Ziviani, editors, *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 146–153. ACM press, New York, NY, USA, 2005.

[2] Jarkko Salojärvi, Kai Puolamäki, and Samuel Kaski. On discriminative joint density modeling. In João Gama, Rui Camacho, Pavel Brazdil, Alípio Jorge, and Luís Torgo, editors, *Machine Learning: ECML 2005*, Lecture Notes in Artificial Intellligence 3720, pages 341–352, Berlin, Germany, 2005. Springer-Verlag.

Chapter 14

Other projects

14.1 Adaptive committee techniques

Matti Aksela, Jorma Laaksonen, Erkki Oja

Combining the results of several classifiers can improve performance because in the outputs of the individual classifiers the errors are not necessarily overlapping. Also the combination method can be adaptive. The two most important features of the member classifiers that affect the committee's performance are their individual error rates and the diversity of the errors. The more different the mistakes made by the classifiers, the more beneficial the combination of the classifiers can be.

Selecting member classifiers is not necessarily simple. Several methods for classifier diversity have been presented to solve this problem. In [1] a scheme weighting similar errors made in an exponential fashion, the Exponential Error Count method, was found to provide good results. Still, the best selection of member classifiers is highly dependent on the combination method used.

We have experimented with several adaptive committee structures. Two effective methods have been the Dynamically Expanding Context (DEC) and Class-Confidence Critic Combining (CCCC) schemes [2]. The DEC algorithm was originally developed for speech recognition purposes. The main idea is to determine just a sufficient amount of context for each individual segment so that all conflicts in classification results can be resolved.

In our CCCC approach the main idea is to try to produce as good as possible an estimate on the classifier's correctness based on its prior behavior for the same character class. This is accomplished by the use of critics that assign a confidence value to each classification. The confidence value is obtained through constructing and updating distribution models of distance values from the classifier for each class in every critic. These distribution models are then used to extract the needed confidence value, based on prior results in addition to the sample being processed. The committee then uses a decision mechanism to produce the final output from the input label information and critic confidence values. The adaptive committee structures have been shown to be able to improve significantly on their members' results [2].

Also classifiers that are adaptive in themselves can be combined using an adaptive committee, and recent experiments have shown that this adaptive combination of adaptive classifiers can produce even better results than either method alone. Combining adaptive classifiers is feasible through the use of a weighting scheme to obtain better balance between the use of prior information and robustness under changing conditions [3].

References

[1] Matti Aksela and Jorma Laaksonen. Using diversity of errors for selecting members of a committee classifier. *Pattern Recognition*, 39(4):608–623, 2006.

[2] Matti Aksela, Ramūnas Girdziušas, Jorma Laaksonen, Erkki Oja, and Jari Kangas. Methods for adaptive combination of classifiers with application to recognition of handwritten characters. *International Journal of Document Analysis and Recognition*, 6(1):23–41, 2003.

[3] Matti Aksela and Jorma Laaksonen. On adaptive confidences for critic-driven classifier combining. In *Proceedings of International Conference on Advances in Pattern Recognition*, volume 2, pages 71–80, 2005.

14.2 Data analysis using the Evolving Tree

Jussi Pakkanen

Modern data analysis problems usually have to deal with very large databases. When the amount of data samples grow to millions or tens of millions, many traditional tools and techniques slow down noticeably. This, combined with the curse of dimensionality, makes problems involving large data sets very difficult to approach.

Our research has focused on finding novel methods to combine neural network systems with large data set manipulation tools of computer science. The goal is to create new neural systems that can be used to analyze huge data bases efficiently while retaining a high precision. We propose the *The Evolving Tree* [1] for this task.

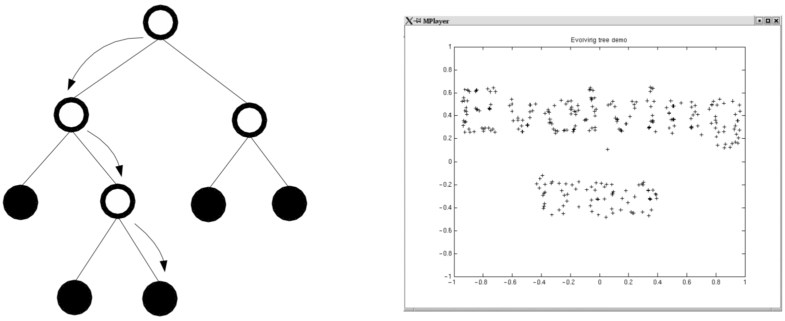


Figure 14.1: The general architecture of the Evolving Tree and an example of adaptation to data.

Figure 14.1 demonstrates the basic properties of the Evolving Tree. The left image shows how the tree is made of two kinds of nodes. The black *leaf nodes* are the actual data analysis nodes, which perform vector coding. The white *trunk nodes* form an efficient search tree to the leaf nodes. The arrows show how a single search on the tree might progress. During training the Evolving Tree grows by creating new leaf nodes to those areas of the data space that are deemed to be underrepresented.

The right image on Figure 14.1 shows how the Evolving Tree adapts to an artificial two-dimensional data set. The dots are the code vectors.. The training had started with a single node, but the tree has grown in size to better explain the data.

Tests on real world industrial data shows that ETree performs favorably when compared to other similar methods [2].

References

[1] Jussi Pakkanen, Jukka Iivarinen and Erkki Oja The Evolving Tree — A Novel Self-Organizing Network for Data Analysis, in *Neural Processing Letters*, Volume 20, Issue 3, pages 199–211, December 2004.

[2] Jussi Pakkanen and Jukka Iivarinen, Analyzing Large Image Databases with the Evolving Tree, in *Proceedings of the Third International Conference on Advances in Pattern Recognition, ICAPR05*, LNCS 3686, pages 192–198, August 2005.

14.3 Independent variable group analysis

Krista Lagus, Antti Honkela, Jeremias Seppä, Paul Wagner

Independent variable group analysis (IVGA) [1] is a principle for grouping observed input variables so that mutual dependences between variables are strong within a group and weak between groups.

In problems with a large number of diverse observations there are often groups of input variables that have strong mutual dependences within the group but which can be considered practically independent of the input variables in other groups. It can be expected that the larger the problem domain, the more independent groups there are. Estimating a model for each independent group separately produces a more compact representation than applying the model to the whole set of variables. Compact representations are computationally beneficial and, moreover, offer better generalization.

Usually such variable grouping is performed by a domain expert, prior to modeling with automatic, adaptive methods. As expert knowledge may be unavailable, or expensive and time-consuming, automating the task can considerably save resources. The IVGA is a practical, efficient and general approach for obtaining compact representations that can be regarded as sparse codes, as well.

The IVGA project is a collaboration with Dr. Esa Alhoniemi (University of Turku) and Dr. Harri Valpola (Helsinki University of Technology, Laboratory of Computational Engineering).

The IVGA algorithm

Any IVGA algorithm consists of two parts, (1) grouping of variables, and (2) construction of a separate model for each variable group. An independent variable grouping is obtained by comparing models with different groupings using a suitable cost function. In principle any model can be used, if the necessary cost function is derived for the model family.

A practical grouping algorithm for implementing the IVGA principle was first presented in [1]. The method used vector quantizers (VQs) learned with variational Bayesian methods [2] to model the individual groups.

Recent development of IVGA [3] has concentrated on extending our algorithmic implementation to handle mixed data consisting of both real valued and nominal variables. A public software package is in preparation and new experiments have been made. In order to allow both real and nominal variables, the vector quantizers were replaced with mixture models so that the mixture components were Gaussians in the real valued case and multinomial distributions in the nominal case.

Experimental results

In this experiment, IVGA was used to group the variables representing attributes of components required by an assembly robot in mounting of the components on a printed circuit board. Finding correct settings for the attributes by hand is difficult and association rules have been applied to model their dependences [4]. Extraction of the rules from the data is computationally heavy, and memory consumption of the data structure (for example, a trie) for the rules is very high for large data sets. Splitting the variables to weakly dependent groups decreases the complexity significantly.

The training data used in this experiment consisted of 24 attributes (17 nominal, 7 real valued) with values for 1000 components. Computation tiems, memory consumption, and prediction accuracy for a separate testing data set using association rules built for the full data set and for the three groups discovered by IVGA are presented in Table 14.1. Splitting of the data using IVGA lead to significant improvements in the efficiency of the obtained model: it accelerated computation of the rules, dramatically reduced the size of the data structure, and decreased the number of the incorrect predictions. On the other hand, the number of missing predictions was clearly larger for the grouped data than for the whole data, because for the first attribute value of every group, no prediction could be made whereas for the whole data, only the prediction for the first attribute could not be obtained [3].

	Whole data	**Grouped data**
Computation time (s)	194	< 1
Size of trie (nodes)	1 054 512	3 914
Correct predictions (%)	38.05	32.45
Incorrect predictions (%)	1.61	0.68
Missing predictions (%)	60.43	66.88

Table 14.1: Summary of the results of the component data experiment. All the quantities for the grouped data are sums over the three groups. Also note that the size of trie is the same as the number of association rules.

In conclusion, the experimental results show that it is worthwhile to group variables according to independence, and that the presented algorithm is able to do this and in doing so, obtains more compact models.

References

[1] K. Lagus, E. Alhoniemi, and H. Valpola, “Independent variable group analysis,” in *International Conference on Artificial Neural Networks - ICANN 2001*, ser. LLNCS, G. Dorffner, H. Bischof, and K. Hornik, Eds., vol. 2130. Vienna, Austria: Springer, August 2001, pp. 203–210.

[2] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, “An introduction to variational methods for graphical models,” in *Learning in Graphical Models*, M. Jordan, Ed. Cambridge, MA, USA: The MIT Press, 1999, pp. 105–161.

[3] K. Lagus, E. Alhoniemi, J. Seppä, A. Honkela, and P. Wagner, “Independent variable group analysis in learning compact representations for data,” in *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR’05)*, T. Honkela, V. Könönen, M. Pöllä, and O. Simula, Eds., Espoo, Finland, June 2005, pp. 49–56.

[4] E. Alhoniemi, T. Knuutila, M. Johnsson, J. Röyhkiö, and O. S. Nevalainen, “Data mining in maintenance of electronic component libraries,” in *Proceedings of the IEEE 4th International Conference on Intelligent Systems Design and Applications*, 2004, vol. 1, pp. 403–408.

14.4 Worldwide research on and using the Self-Organizing Map

Teuvo Kohonen, Timo Honkela and Matti Pöllä

The Self-Organizing Map [2] (SOM) is an effective method for the analysis and visualization of high-dimensional data. It converts complex, nonlinear statistical relationships between high-dimensional data items into simple geometric relationships on a low-dimensional display. As it thereby compresses information while preserving the most important topological and metric relationships of the primary data items on the display, it may also be thought to produce some kind of abstractions. These two aspects, visualization and abstraction, can be utilized in a number of ways in complex tasks.

The Self-Organizing Map has attracted a great deal of interest among researches and practitioners in a wide variety of fields. The SOM has been analyzed extensively, a number of variants have been developed and, perhaps most notably, it has been applied extensively within fields ranging from engineering sciences to medicine, biology, and economics. Comprehensive lists of scientific papers that use the algorithms, have benefited from them, or contain analyses of them has earlier been collected covering the time until early 2002 [3, 4]. A new, for the moment unpublished update of the bibliography is under preparation (by M. Pöllä, T. Honkela and T. Kohonen). Based on this newest collection, we can report that by the end of 2005, there were altogether approximately 7000 references.

The first two scientific events dedicated to the SOM, its theory and applications, the Workshop on Self-Organizing Maps, were organized at the Helsinki University of Technology in 1997 and 1999. Since then, the WSOM has been organized biannually (in 2001 in Lincoln, England, in 2003 in Hibikino, Kitakyushu, Japan and in 2005 in Paris by University Paris 1). In addition, the SOM is often a specific theme in conferences related both to computer science and to the several application areas of the SOM ranging from industry to bioinformatics.

In summary, the research on the Self-Organizing map is very active still more than 20 years after its original invention and publication [1]. New results on its theory, extensions and applications are published in practice on daily basis. The practical importance of the SOM can also be highlighted by the fact that the Self-Organizing Map has been included in a large number of commercial analytical software packages both by large companies producing general purpose tools or by companies dedicated to the development of SOM-based tools and applications .

In the following, we describe a novel development, a model called Self-Organizing Neural Projection and give an overview on the body of research in our laboratory based on the Self-Organizing Map.

References

[1] T. Kohonen (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43:59-69.

[2] T. Kohonen (2001). *Self-Organizing Maps.* Third, extended edition. Springer.

[3] S. Kaski, J. Kangas, T. Kohonen (1998). Bibliography of self-organizing map (SOM) papers: 1981-1997. *Neural Computing Surveys*, 1: 102-350.

[4] M. Oja, S. Kaski, and T. Kohonen (2003). Bibliography of Self-Organizing Map (SOM) Papers: 1998-2001 Addendum. *Neural Computing Surveys*, 3: 1-156.

14.5 Self-Organizing Neural Projections

Teuvo Kohonen

The SOM algorithm was developed for the creation of abstract-feature maps [3]. It has been accepted widely as a data-mining tool, and the principle underlying it may also explain how the feature maps of the brain are formed [4]. However, it is not correct to use this algorithm for a model of pointwise neural projections such as the somatotopic maps or the maps of the visual field, first of all, because the SOM does not transfer signal patterns: the winner-take-all function at its output only defines a singular response. Neither can the original SOM produce superimposed responses to superimposed stimulus patterns.

The SOM was not the first model of cortical organization (cf., e.g., the line detector model of v.d. Malsburg [6], and Amari's "synaptic field" model of the Type 1 pointwise maps [1]. Unfortunately, none of these attempts was a success. For instance, Amari's maps were not globally ordered. They were always parcelled into small, ordered patches, between which the ordering changed abruptly. On the other hand, v.d. Malsburg's model was "brittle," because the reported ordering only took place for a parameter value that was defined by three decimal places, and the maps could not be generalized.

The reason for the failure of the earlier models was that they were solely based on excitatory and inhibitory lateral connections, and the Hebbian rule of synaptic plasticity. By means of the lateral connections, the output activity was first clustered spatially, and the adaptation then took place in these clusters in proportion to input and output activities. However, when using these models, the activity clusters, in order to obtain globally ordered maps, should have been very wide, of the same order of magnitude as the dimension of the array. But then one could have hardly regarded such wide activity clusters as pointwise output responses. On the other hand, if the clusters had been made smaller, the maps would only have been organized into small, disjoint local patches. Also, such clustering would have disturbed the transfer of the original signals.

In the biological realms, genetic information defines a very rough initial order of the neural projections. Refinement of this order begins already prenatally, by means of endogenous signals generated by the network itself. The final resolution of the mapping, and optimization of the neural resources (magnification factor), however, are achieved postnatally, according to the sensory experiences. It has been demonstrated that exposing newborn rats to continuous moderate-level acoustic noise, the development and refinement of the tonotopic maps will be delayed long beyond normal periods [2]. It has also been shown that the exposure of infant rats to complex tone sequences results in altered auditory cortex organization [7]. These observations prove that the input-driven organization of the brain maps is a fact and needs a new theoretical model.

We have recently introduced a novel self-organizing system model related to the SOM that has a linear transfer function for patterns and combinations of patterns all the time [5]. Starting from a randomly interconnected pair of neural layers, and using random mixtures of patterns for training, it creates a pointwise-ordered projection from the input layer to the output layer. If the input layer consists of feature detectors, the output layer forms a feature map of the inputs. More detailed description of the model can be found in [5].

References

[1] Amari, S.-I. (1980). Topographic organization of nerve fields. *Bulletin of Mathematical Biology*, 42:339–364.

[2] Chang, E.F. and Merzenich, M.M. (2003). Environmental Noise Retards Auditory Cortical Development *Science*, 300 (5618): 498–502.

[3] T. Kohonen (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43:59-69.

[4] Kohonen, T. (2001). *Self-Organizing Maps*. Third, extended edition. Springer.

[5] Kohonen, T. (2005). Pointwise organizing projections. M. Cottrell (ed.), *Proceedings of the 5th Workshop on Self-Organizing Maps*, Paris, pp. 1-8.

[6] von der Malsburg, Ch. (1973). Self-organization of orientation sensitive cells in the striate cortex. *Kybernetik*, 14, 85–100.

[7] Nakahara, H. Zhang, L.I. and Merzenich, H.H. (2004). Specialization of primary auditory cortex processing by sound exposure in the "critical period". PNAS 101: 7170–7174

14.6 Applications of the Self-Organizing Map

Teuvo Kohonen, Olli Simula, Erkki Oja, Samuel Kaski, Timo Honkela, Jukka Iivarinen, Markus Koskela, Mikko Kurimo, Jorma Laaksonen, Krista Lagus, Kimmo Raivio, Miki Sirola

The Self-Organizing Map is applied in several areas of research in our laboratory. The application areas incluce:

- content-based image retrieval (see Chapter 7.1 for a detailed description),
- modeling the emergence of cognitive and conceptual representations (Chapter 8.2),
- language modeling (Chapters 9.3 and 10.3) and word sense disambiguation using document maps (Chapter 10.2), and
- multiple applications in the area of intelligent data engineering including the analysis of mobile radio access network (Chapter 11.2), decision support (Chapter 11.4), and modeling dependencies in data (Chapter 11.5).

In addition, a new hierarchical self-organizing system called the Evolving Tree has been described in Chapter 14.2.

Publications of the Neural Networks Research Centre

Publications are in alphabetical order by the first author.

2004

1. Beliaev, I.; Kozma, R.; Lendasse, A. Robust Time Series Prediction Using KIII Model. *Proc. of IDS'04 Symposium, FedEx Institute of Technology (FIT)*, University of Memphis, TN, USA, April 24-26, 2004.

2. Bochko, V.; Kalenova, D.; Harva, M.; Parkkinen, J. Spectral Color Picking Technique Using Nonlinear PCA. Lappeenranta, Finland: Lappeenranta University of Technology, 2004. (Technical Report).

3. Creutz, M.; Lagus, K. Induction of a Simple Morphology for Highly-Inflecting Languages. *Proc. of 7th Meeting of the ACL Special Interest Group in Computational Phonology (SIGPHON)*, Barcelona, Spain, July 26, 2004. pp. 43-51.

4. Creutz, M.; Lindén, K. Morpheme Segmentation Gold Standards for Finnish and English. Espoo, Finland: 2004. (Helsinki University of Technology, Publications in Computer and Information Science Report A77).

5. Girdziusas, R.; Laaksonen, J. Gaussian Process Regression with Fluid Hyperpriors. *Proc. of 11th International Conference on Neural Information Processing*, Calcutta, India, November 22-25, 2004. Berlin Heidelberg 2004, Springer-Verlag, pp. 567-572.

6. Harva, M.; Kaban, A. Bayesian Inference of Independent Components from Elliptical Stellar Population Spectra. Birmingham, U.K.: School of Computer Science, The University of Birmingham, 2004. (Technical Report).

7. Hiisilä, H.; Bingham, E. Dependencies Between Tanscription Factor Binding Sites: Comparison Between ICA, NMF, PLSA and Frequent Sets. *Proc. of 4th IEEE International Conference on Data Mining*, Brighton, UK, November 1-4, 2004. pp. 114-121.

8. Himberg, J.; Hyvärinen, A.; Esposito, F. Validating the Independent Components of Neuroimaging Time-Series via Clustering and Visualization. *Neuroimage*, 2004. Vol. 22, No. 3, pp. 1214-1222.

9. Hirsimäki, T.; Kurimo, M. Decoder Issues in Unlimited Finnish Speech Recognition. *Proc. of 6th Nordic Signal Processing Symposium (Norsig 2004)*, Espoo, Finland, June 9-11, 2004. pp. 320-323.

10. Honkela, A. Approximating Nonlinear Transformations of Probability Distributions for Nonlinear Independent Component Analysis. *Proc. of 2004 IEEE International Joint Conference on Neural Networks (IJCNN 2004)*, Budapest, Hungary, July 25-29, 2004. pp. 2169-2174.

11. Honkela, A.; Harmeling, S.; Lundqvist, L.; Valpola, H. Using Kernel PCA for Initialisation of Variational Bayesian Nonlinear Blind Source Separation Method. *Proc. of Fifth International Conference Independent Component Analysis and Blind Signal Separation (ICA 2004)*, Granada, Spain, September 22-24, 2004. Springer-Verlag, pp. 790-797.

12. Honkela, A.; Valpola, H. Variational Learning and Bits-Back Coding: An Information-Theoretic View to Bayesian Learning. *IEEE Transactions on Neural Networks*, 2004. Vol. 15, No. 4, pp. 800-810.

13. Honkela, T.; Hyvärinen, A. Linguistic Feature Extraction using Independent Component Analysis. *Proc. of International Joint Conference on Neural Networks, IJCNN 2004*, Budapest, Hungary, July 25-29, 2004. pp. 279-284.

14. Honkela, T.; Nordfors, R.; Tuuli, R. Document Maps for Competence Management. *Proc. of Symposium on Professional Practice in AI, IFIP 18th World Computer Congress*, Toulouse, France, August 22-27, 2004. pp. 31-39.

15. Härmä, A.; Somervuo, P. Classification of the Harmonic Structure in Bird Vocalization. *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'04)*, Montreal, Canada, May 17-21, 2004. Vol. V, pp. 701-704.

16. Iivarinen, J.; Pakkanen, J.; Rauhamaa, J. SOM-based System for Web Surface Inspection. *Proc. SPIE 5303*, Machine Vision Applications in Industrial Inspection XII, San Jose, California, January 21-22, 2004. pp. 178-187.

17. Iivarinen, J.; Rautkorpi, R.; Pakkanen, J.; Rauhamaa, J. Content-Based Retrieval of Surface Defect Images with PicSOM. *International Journal of Fuzzy Systems*, 2004. Vol. 6, No. 3, pp. 160-166.

18. Ilin, A.; Achard, S.; Jutten, S. Bayesian versus Constrained Structure Approaches for Source Separation in Post-Nonlinear Mixtures. *Proc. of International Joint Conference on Neural Networks (IJCNN 2004)*, Budapest, Hungary, July 25-29, 2004. pp. 2181-2186.

19. Ilin, A.; Honkela, A. Post-Nonlinear Independent Component Analysis by Variational Bayesian Learning. *Proc. of 5th Int. Symp. on Independent Component Analysis and Blind Signal Separation (ICA2004)*, Granada, Spain, September 22-24, 2004. pp. 766-773.

20. Ilin, A.; Valpola, H.; Oja, E. Nonlinear Dynamical Factor Analysis for State Change Detection. *IEEE Transactions on Neural Networks*, 2004. Vol. 15, No. 3, pp. 559-575.

21. Inki, M. A Model for Analyzing Dependencies Between Two ICA Features in Natural Images. *Proc. of Fifth International Conference on Independent Component Analysis (ICA2004)*, Granada, Spain, Sept. 22-24, 2004. pp. 914-921.

22. Inki, M. Natural Image Patch Statistics Conditioned on Acitivity of an Independent Component. Espoo, Finland: Helsinki University of Technology, 2004. (Publications in Computer and Information Science Report A79).

23. Jutten, C.; Karhunen, J. Advances in Blind Source Separation (BSS) and Independent Component Analysis (ICA) for Nonlinear Mixtures. *International Journal of Neural Systems*, 2004. Vol. 14, No. 5, pp. 267-292.

24. Jørgensen, A.C.; Rantanen, J.; Luukkonen, P.; Laine, S.; Yliruusi, J. Visualization of a Pharmaceutical Unit Operation: Wet Granulation. *Analytical Chemistry*, 2004. Vol. 76, No. 18, pp. 5331-5338.

25. Kaban, A.; Bingham, E.; Hirsimäki, T. Learning to Read Between the Lines: The Aspect Bernoulli Model. *Proc. of 4th SIAM International Conference on Data Mining*, Lake Buena Vista, Florida, USA, April 22-24, 2004. pp. 462-466.

26. Karp, E.; Gävert, H.; Särelä, J.; Vigário, R. Independent Component Analysis Decomposition of Structural MRI. *Proc. of 2nd IASTED International Conference on Biomedical Engineering (BioMED 04)*, Innsbruck, Austria, Feb. 16-18, 2004. pp. 83-87.

27. Karp, E.; Vigário, R. Unsupervised MRI Tissue Classification by Support Vector Machines. *Proc. of 2nd IASTED International Conference on Biomedical Engineering (BioMED 04)*, Innsbruck, Austria, Feb. 16-18, 2004. pp. 88-91.

28. Kaski, S.; Sinkkonen, J. Principle of Learning Metrics for Data Analysis. *Journal of VLSI Signal Processing*, Special Issue on Machine Learning for Signal Processing, 2004. Vol. 37, pp. 177-188.

29. Kohonen, T. Self-Organizing Tactile Maps. Espoo, Finland: Helsinki University of Technology, 2004. (Publications in Computer and Information Science, Report A76).

30. Koskela, M.; Laaksonen, J.; Oja, E. Entropy-Based Measures for Clustering and SOM Topology Preservation Applied to Content-Based Image Indexing and Retrieval. *Proc. of 17th International Conference on Pattern Recognition (ICPR 2004)*, Cambridge, UK, August 2004. pp. 1005-1008.

31. Koskela, M.; Laaksonen, J.; Oja, E. Use of Image Subset Features in Image Retrieval with Self-Organizing Maps. *Proc. of 3rd International Conference on Image and Video Retrieval (CIVR 2004)*, Dublin, Ireland, July 2004. pp. 508-516.

32. Kurimo, M.; Turunen, V.; Ekman, I. An Evaluation of a Spoken Document Retrieval Baseline System in Finnish. *Proc. of International Conference on Spoken Language Processing (ICSLP 2004*, Jeju Island, Korea, Oct. 4-8, 2004. pp. II-1585-1588.

33. Kurimo, M.; Turunen, V.; Ekman, I. Speech Transcription and Spoken Document Retrieval in Finnish. *Proc. of Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI 2004)*, Martigny, Switzerland, June 21-23, 2004.

34. Kurimo, M.; Zhou, B.; Huang, R.; Hansen, J. H. L. Language Modeling Structures in Audio Transcription for Retrieval of Historical Speeches. *Proc. of European Signal Processing Conference, EUSIPCO 2004*, Vienna, Austria, Sept. 6-10, 2004. pp. 557-560.

35. Kylväjä, M.; Hätönen, K.; Kumpulainen, P.; Laiho, J.; Lehtimäki, P.; Raivio, K.; Vehviläinen, P. Trial Report on Self-Organizing Map based Analysis Tool for Radio Networks. *Proc. of IEEE Semiannual Vehicular Technology Conference*, Milan, Italy, May 17-19, 2004.

36. Könönen, V. Asymmetric Multiagent Reinforcement Learning. *Web Intelligence and Agent Systems: An International Journal (WIAS)*, 2004. Vol. 2, No. 2, pp. 105-121.

37. Könönen, V. Gradient Descent for Symmetric and Asymmetric Multiagent Reinforcement Learning. Espoo, Finland: 2004. 18 p. (Helsinki University of Technology, Publications in Computer and Information Science, Report A78).

38. Könönen, V. Hybrid Model for Multiagent Reinforcement Learning. *Proc. of International Joint Conference on Neural Networks (IJCNN-2004)*, Budapest, Hungary, July 25-29, 2004. pp. 1793-1798.

39. Könönen, V. Policy Gradient Method for Team Markov Games. *Proc. of Fifth International Conference on Intelligent Data Engineering and Automated Learning (IDEAL-2004)*, Exeter, UK, August 25-27, 2004. pp. 733-739.

40. Könönen, V.; Oja, E. Asymmetric Multiagent Reinforcement Learning in Pricing Applications. *Proc. of International Joint Conference on Neural Networks (IJCNN-2004)*, Budapest, Hungary, July 25-29, 2004. pp. 1097-1102.

41. Laakso, M.-L.; Leinonen, L.; Lindblom, N.; Joutsiniemi, S.-L.; Kaski, M. Wrist Actigraphy in Estimation of Sleep and Wake in Intellectually Disabled Subjects with Motor Handicaps. *Sleep Medicine*, 2004. Vol. 5, pp. 541-550.

42. Laaksonen, J.; Koskela, M.; Oja, E. Class Distributions on SOM Surfaces for Feature Extraction and Object Retrieval. *Neural Networks*, 2004. Vol. 17, No. 8-9, pp. 1121-1133.

43. Lagus, K.; Kaski, S.; Kohonen, T. Mining Massive Document Collections by the WEBSOM Method. *Information Sciences*, 2004. Vol. 163, No. 1-3, pp. 135-156.

44. Laine, S.; Similä, T. Using SOM-Based Data Binning to Support Supervised Variable Selection. *Proc. of 11th International Conference on Neural Information Processing (ICONIP 2004)*, Calcutta, India, Nov. 22-25, 2004. pp. 172-180.

45. Lendasse, A.; François, D.; Rossi, F.; Wertz, V.; Verleysen, M. Mutual Information for the Selection of Relevant Variables in Spectrometric Nonlinear Modeling. *Proc. of Chimiométrie 2004*, Paris, France, November 30 - December 1, 2004.

46. Lendasse, A.; Oja, E.; Simula, O.; Verleysen, M. Time Series Benchmarking Competition: The CATS Benchmark. *Proc. of International Joint Conference on Neural Networks (IJCNN'04)*, Budapest, Hungary, July 26-29, 2004. pp. 1615-1620.

47. Lendasse, A.; Simon, G.; Wertz, V.; Verleysen, M. Fast Bootstrap for Least-Square Support Vector Machines. *Proc. of European Symposium on Artificial Neural Networks (ESANN 2004)*, Bruges, Belgium, April 28-30, 2004. pp. 525-530.

48. Lendasse, A.; Wertz, V.; Simon, G.; Verleysen, M. Fast Bootstrap Applied to LS-SVM for Long Term Prediction of Time Series. *Proc. of International Joint Conference on Neural Networks (IJCNN 2004)*, Budapest, Hungary, July 25-29, 2004. pp. 705-710.

49. Lindén, K. Evaluation of Linguistic Features for Word Sense Disambiguation with Self-Organized Document Maps. *Journal of the Computers and the Humanities*, 2004. Vol. 38, No. 4, pp. 417-435.

50. Luyssaert, S.; Sulkava, M.; Raitio, H.; Hollmén, J. Evaluation of Forest Nutrition Based on Large-Scale Foliar Surveys: Are Nutrition Profiles the Way of the Future? *Journal of Environmental Monitoring*, 2004. Vol. 6, No. 2, pp. 160-167.

51. Müller, K.-R.; Vigário, R.; Meinecke, F.; Ziehe, A. Blind Source Separation Techniques for Decomposing Event Related Brain Signals. *Int. Journal of Bifurcation and Chaos*, 2004. Vol. 14, No. 2, pp. 773-792 .

52. Mäntyjärvi, J.; Himberg, J.; Kangas, P.; Tuomela, U.; Huuskonen, P. Sensor Signal Data Set for Exploring Context Recognition of Mobile Devices. *Proc. of Workshop "Benchmarks and a database for context recognition"*, in conjuction with the 2nd Int. Conf. on Pervasive Computing (PERVASIVE 2004), Linz/Vienna, Austria, April 18-23, 2004. (Electronic publication, Swiss Federal Institute of Technology Zurich, Electronics laboratory)

53. Mäntyjärvi, J.; Nybergh, K.; Himberg, J.; Hjelt, K. Touch Detection System for Mobile Terminals. *Proc. of Mobile Human-Computer Interaction - MobileHCI 2004: 6th International Symposium*, Glasgow, UK, September 13-16, 2004. Heidelberg 2004, Springer-Verlag, pp. 331-336.

54. Nikkilä, J.; Roos, C.; Kaski, S. Exploring Dependencies Between Yeast Stress Genes and their Regulators. *Proc. of International Conference on Intelligent Data Engineering and Automated Learning (IDEAL 2004)*, Exeter, UK, August 25-27, 2004. Springer, pp. 92-98.

55. Oja, E. Applications of Independent Component Analysis. *Proc. of International Conference on Neural Information Processing (ICONIP'04)*, Calcutta, India, Nov. 22-25, 2004.

56. Oja, E. Blind Source Separation: Neural Net Principles and Applications. *Proc. of SPIE Defense and Security Symposium*, Orlando, USA, April 12-16, 2004. Vol. 5439, pp. 1-14.

57. Oja, E. Finding Clusters and Components by Unsupervised Learning. *Proc. of IAPR International Workshop on Statistical Pattern Recognition (SPR2004)*, Lisbon, Portugal, August 18-20, 2004. pp. 1-15.

58. Oja, E. Patterns, Clusters, and Components - What Data is Made of. *Proc. of International Joint Conference on Neural Networks (IJCNN'04)*, Budapest, Hungary, July 26-29, 2004. p. 3.

59. Oja, E.; Harmeling, S; Almeida, L. Independent Component Analysis and Beyond. *Signal Processing*, 2004. Vol. 84, No. 2, pp. 215-216.

60. Oja, E.; Plumbley, M. Blind Separation of Positive Sources by Globally Convergent Gradient Search. *Neural Computation*, 2004. Vol. 16, No. 9, pp. 1811-1825.

61. Oja, M.; Sperber, G.; Blomberg, J.; Kaski, S. Grouping and Visualizing Human Endogenous Retroviruses by Bootstrapping Median Self-organizing Maps. *Proc. of IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, San Diego, California, USA, October 7-8, 2004. pp. 95-101.

62. Ojala, T.; Koskela, M.; Matinmikko, E.; Rautiainen, M.; Laaksonen, J.; Oja, E. Task-Based User Evaluation of Content-Based Image Database Browsing Systems. *Proc. of 3rd International Conference on Image and Video Retrieval (CIVR 2004)*, Dublin, Ireland, July 2004. pp. 234-242.

63. Pakkanen, J.; Iivarinen, J. A Novel Self-Organizing Neural Network for Defect Image Classification. *Proc. of International Joint Conference on Neural Networks*, Budapest, Hungary, July 25-29, 2004. pp. 2553-2558.

64. Pakkanen, J.; Iivarinen, J.; Oja, E. The Evolving Tree - a Novel Self-Organizing Network for Data Analysis. *Neural Processing Letters*, 2004. Vol. 20, No. 3, pp. 199-211.

65. Pakkanen, J.; Turkulainen, P. The Evolving Tree Software Package. 2004. http://www.cis.hut.fi/research/etree/

66. Peltonen, J.; Klami, A.; Kaski, S. Improved Learning of Riemannian Metrics for Exploratory Data Analysis. *Neural Networks*, 2004. Vol. 17, pp. 1087-1100.

67. Peltonen, J.; Sinkkonen, J.; Kaski, S. Sequential Information Bottleneck for Finite Data. *Proc. of Twenty-First International Conference on Machine Learning (ICML 2004)*, Banff, Alberta, Canada, July 4-8, 2004. pp. 647-654.

68. Plumbley, M.; Oja, E. A "Non-negative PCA" Algorithm for Independent Component Analysis. *IEEE Trans. on Neural Networks*, 2004. Vol. 15, No. 1, pp. 66-76.

69. Principe, J.; Oja, E.; Xu, L.; Cichocki, A.; Erdogmus, D. Guest Editorial - Special issue on information theoretic learning. *IEEE Trans. on Neural Networks*, 2004. Vol. 15, No. 4, pp. 789-791.

70. Puolamäki, K.,; Savia, E.; Sinkkonen, J.; Kaski, S. Two-Way Latent Grouping Model for User Preference Prediction. Espoo, Finland: Helsinki University of Technology, 2004. (Publications in Computer and Information Science, Report A80).

71. Pylkkönen, J.; Kurimo, M. Duration Modeling Techniques for Continuous Speech Recognition. *Proc. of 8th International Conference on Spoken Language Processing (Interspeech 2004)*, Jeju Island, Korea, October 4-8, 2004. pp. 385-388.

72. Pylkkönen, J.; Kurimo, M. Using Phone Durations in Finnish Large Vocabulary Continuous Speech Recognition. *Proc. of 6th Nordic Signal Processing Symposium (Norsig 2004)*, Espoo, Finland, June 9-11, 2004. pp. 324-327.

73. Raiko, T. Partially Observed Values. *Proc. of International Joint Conference on Neural Networks (IJCNN 2004)*, Budapest, Hungary, July 25-29, 2004. pp. 2825-2830.

74. Raiko, T. The Go-Playing Program Called Go81. *Proc. of Finnish Artificial Intelligence Conference (STeP 2004)*, Helsinki, Finland, September 1-3, 2004. pp. 197-206.

75. Raitio, J.; Vigário, R.; Särelä, J.; Honkela, T. Assessing Similarity of Emergent Representations Based on Unsupervised Learning. *Proc. of International Joint Conference on Neural Networks (IJCNN 2004)*, Budapest, Hungary, July 25-29, 2004. pp. 597-602.

76. Raju, K.; Phani Sudheer, B. Blind Source Separation for Interference Cancellation - A Comparision of Several Spatial and Temporal Statistics Based Techniques. *Proc. of 3rd Int. Workshop on the Internet, Telecommunications and Signal Processing*, Adelaide, Australia, Dec. 20-22, 2004.

77. Raju, K.; Ristaniemi, T.; Karhunen, J. Semi-Blind Interference Suppression on Coherent Multipath Environments. *Proc. of First Int. Symposium on Control Communications and Signal Processing (ISCCSP 2004)*, Hammamet, Tunisia, March 21-24, 2004. pp. 283-286.

78. Raju, K.; Särelä, J. A Denoising Source Separation Based Approach to Interference Cancellation for DS-CDMA Array Systems. *Proc. of 38th Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, Nov. 07-10, 2004. pp. 1111-1114.

79. Rautkorpi, R.; Iivarinen, J. A Novel Shape Feature for Image Classification and Retrieval. *Proc. of International Conference on Image Analysis and Recognition*, Porto, Portugal, September 29-October 1, 2004. pp. 753-760.

80. Rautkorpi, R.; Iivarinen, J. Content-Based Image Retrieval of Web Surface Defects with PicSOM. *Proc. of International Joint Conference on Neural Networks*, Budapest, Hungary, July 25-29, 2004. pp. 1863-1868.

81. Rytkönen, K.; Valpola, H.; Särelä, J. DSS Matlab Package, Software Package. 2004. http://www.cis.hut.fi/projects/dss/

82. Salojärvi, J.; Puolamäki, K.; Kaski, S. Relevance Feedback from Eye Movements for Proactive Information Retrieval. *Proc. of Workshop on Processing Sensory Information for Proactive Systems (PSIPS 2004)*, Oulu, Finland, June 14-15, 2004. pp. 37-42.

83. Savia, E.; Kaski, S.; Tuulos, V.; Myllymäki, P. On Text-Based Estimation of Document Relevance. *International Joint Conference on Neural Networks 2004*, Budapest, Hungary, July 25-29, 2004. pp. 3275-3280.

84. Simon, G.; Lendasse, A.; Cottrell, M.; Fort, J.-C.; Verleysen, M. Double Quantization of the Regressor Space for Long-Term Time Series Prediction: Method and Proof of Stability. *Neural Networks*, 2004. No. 8-9, pp. 1169-1181.

85. Sinkkonen, J.; Nikkilä, J.; Lahti, L.; Kaski, S. Associative Clustering. *Proc. of 15th European Conference on Machine Learning (ECML 2004)*, Pisa, Italy, Sept. 20-24, 2004. pp. 396-406.

86. Sirola, M. Decision Concepts. *International Scientific Journal of Computing*, 2004. Vol. 3, No. 2, pp. 18-22.

87. Sirola, M.; Lampi, G.; Parviainen, J. Neuro Computing in Knowledge-Based Decision Support Systems. *Proc. of EHPG-Meeting of OECD Halden Reactor Project*, Sandefjord, Norway, May 9-14, 2004. 9 p.

88. Sirola, M.; Lampi, G.; Parviainen, J. Using Self-Organizing Map in a Computerized Decision Support System. *Proc. of International Conference on Neural Information Processing (ICONIP)*, Calcutta, India, November 22-25, 2004.

89. Skripal, P.; Honkela, T. Framework for Modeling Emotions in Communities of Agents. *Proc. of 11th Finnish Artificial Intelligence Conference*, Vantaa, Finland, September 1-3, 2004. pp. 163-172.

90. Somervuo, P. Comparison of ML, MAP, and VB Based Acoustic Models in Large Vocabulary Speech Recognition. *Proc. of 8th International Conference on Spoken Language Processing (ICSLP)*, Jeju Island, Korea, October 5-8, 2004. Vol. I, pp. 701-704.

91. Somervuo, P. Online Algorithm for the Self-Organizing Map of Symbol Strings. *Neural Networks*, 2004. Vol. 17, pp. 1231-1239.

92. Somervuo, P.; Härmä, A. Bird Song Recognition Based on Syllable Histograms. *Proc.of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'04)*, Montreal, Canada, May 17-21, 2004. Vol. V, pp. 825-828.

93. Sorjamaa, A.; Lendasse, A.; François, D.; Verleysen, M. Business Plans Classification with Locally Pruned Lazy Learning Models. *Proc. of ACSEG 2004, Connectionist Approaches in Economics and Management Sciences*, Lille, France, November 18-19, 2004. pp. 112-119.

94. Sulkava, M.; Tikka, J.; Hollmén, J. Sparse Regression for Analyzing the Development of Foliar Nutrient Concentrations in Coniferous Trees. *Proc. of Fourth International Workshop on Environmental Applications of Machine Learning (EAML 2004)*, Bled, Slovenia, September 2004. pp. 57-58.

95. Tikka, J.; Hollmén, J. Learning Linear Dependency Trees from Multivariate Time-series Data. *Proc. of Workshop on Temporal Data Mining: Algorithms, Theory and Applications*, in conjunction with The Fourth IEEE International Conference on Data Mining, Brighton, UK, November 2004.

96. Valpola, H., Harva, M. & Karhunen, J. Hierarchical Models of Variance Sources. *Signal Processing*, 2004. Vol. 84, No. 2, pp. 267-282.

97. Valpola, H.; Särelä, J. Accurate, Fast and Stable Denoising Source Separation Algorithms. *Proc. of 5th International Conference on Independent Component Analysis and Blind Signal Separation (ICA 2004)*, Granada, Spain, Sept. 22-24, 2004. pp. 65-72.

98. Vesanto, J. & Hollmén, J. An Automated Report Generation Tool for the Data Understanding Phase. In: Abraham, A. , Jain, L. & van der Zwaag, B. J. (eds.), *Innovations in Intelligent Systems: Design, Management and Applications, Studies in Fuzziness and Soft Computing Vol. 140.* Heidelberg 2004, Springer (Physica) Verlag, chapter 5.

99. Väyrynen, J. J.; Honkela, T. Word Category Maps based on Emergent Features Created by ICA. *Proc. of 11th Finnish Artificial Intelligence Conference*, Vantaa, Finland, September 1-3, 2004. pp. 173-185.

100. Väyrynen, J. J.; Honkela, T.; Hyvärinen, A. Independent Component Analysis of Word Contexts and Comparison with Traditional Categories. *Proc. of 6th Nordic Signal Processing Symposium - NORSIG 2004*, Espoo, Finland, June 9-11, 2004. pp. 300-303. (Electronic publication, http://wooster.hut.fi/publications/norsig2004/)

101. Ylipaavalniemi, J.; Vigário, R. Analysis of Auditory fMRI Recordings via ICA: A Study on Consistency. *Proc. of International Joint Conference on Neural Networks (IJCNN 2004)*, Budapest, Hungary, July 25-29, 2004. pp. 249-254.

102. Yuan, Z.; Oja, E. A FastICA Algorithm for Non-Negative Independent Component Analysis. *Proc. of 5th International Symposium on Independent Component Analysis and Blind Source Separation*, Granada, Spain, Sep. 22-24, 2004. pp. 1-8.

2005

1. Aksela, M.; Laaksonen, J. On Adaptive Confidences for Critic-Driven Classifier Combining. *Proc. of 3rd International Conference on Advances in Pattern Recognition 2005*, Bath, United Kingdom, August 22-25, 2005. pp. 71-80.

2. Bas, P. A Quantization Watermarking Technique Robust to Linear and Non-Linear Valumetric Distortions Using a Fractal Set of Foating Quantizers. *Proc. of Information Hiding Workshop 2005*, Barcelona, Spain, June 6-8, 2005. pp. 106-117.

3. Bas, P.; Hurri, J. Security of DM Quantization Watermarking Scheme: A Practical Study for Digital Images. *Proc. of International Workshop of Digital Watermarking*, Sienna, Italy, September 15-17, 2005. pp. 186-200.

4. Borisov, S.; Ilin, A.; Vigário, R.; Kaplan, A. Source Localization of Low- and High-Amplitude Alpha Activity: A Segmental and DSS Analysis. *Proc. of 11th Annual Meeting of Organization for Human Brain Mapping*, Toronto, Canada, June 12-16, 2005; Neuroimage, V. 26., Suppl. 1. p. 38.

5. Bounsaythip, C.; Hollmén, J.; Kaski, S.; Oresic, M. (eds.) *Proceedings of KRBIO05, Symposium on Knowledge Representation in Bioinformatics*, Espoo, Finland, June 15-17, 2005. Espoo, Finland 2005, Helsinki University of Technology. 50 p.

6. Broman, S.; Kurimo, M. Methods for Combining Language Models in Speech Recognition. *Proc. of 9th European Conference on Speech Communication and Technology, Interspeech 2005*, Lisbon, Portugal, September 4-8, 2005. pp. 1317-1320.

7. Corona, F.; Lendasse, A. Input Selection and Function Approximation using the SOM: an Application to Spectrometric Modeling. *Proc. of 5th Workshop on Self-Organizing Maps (WSOM'05)*, Paris, France, September 5-8, 2005. pp. 653-660.

8. Creutz, M.; Lagus, K. Inducing the Morphological Lexicon of a Natural Language from Unannotated Text. *Proc. of International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05)*, Espoo, Finland, June 15-17, 2005. pp. 106-113.

9. Creutz, M.; Lagus, K. Unsupervised Morpheme Segmentation and Morphology Induction from Text Corpora Using Morfessor 1.0. Espoo, Finland: Helsinki University of Technology, 2005. (Publications in Computer and Information Science, Report A81).

10. Creutz, M.; Lagus, K.; Lindén, K.; Virpioja, S. Morfessor and Hutmegs: Unsupervised Morpheme Segmentation for Highly-Inflecting and Compounding Languages. *Proc. of Second Baltic Conference on Human Language Technologies*, Tallinn, Estonia, April 4-5, 2005. pp. 107-112.

11. Duch, W.; Kacprzyk, J.; Oja, E.; Zadrozny, S. (eds.) *Artifical Neural Networks: Biological Inspirations - ICANN 2005.* Lecture Notes in Computer Science 3696. Berlin, Germany 2005, Springer.

12. Duch, W.; Kacprzyk, J.; Oja, E.; Zadrozny, S. (eds.) *Artifical Neural Networks: Formal Models and Their Applications - ICANN 2005.* Lecture Notes in Computer Science 3697. Berlin, Germany 2005, Springer.

13. Elo, L. L.; Lahti, L.; Skottman, H.; Kyläniemi, M.; Lahesmaa, R.; Aittokallio, T. Integrating Probe-Level Expression Changes Across Generations of Affymetrix Arrays. *Nucleic Acids Research*, 2005. Vol. 33, No. 22, pp. e193.

14. Esposito, F.; Scarabino, T.; Hyvärinen, A.; Himberg, J.; Formisano, E.; Comani, S.; Tedeschi, G.; Goebel, R.; Seifritz, E.; Di Salle, F. Independent Component Analysis of fMRI Group Studies by Self-Organizing Clustering. *Neuroimage*, 2005. Vol. 25, No. 1, pp. 193-205.

15. Girdziusas, R.; Laaksonen, J. Gaussian Processes of Nonlinear Diffusion Filtering. *Proc. of IEEE-INNS-ENNS Int. Joint Conference on Neural Networks*, Montréal, Canada, July 31 - August 4, 2005. pp. 1012-1017.

16. Girdziusas, R.; Laaksonen, J. Jacobi Alternative to Bayesian Evidence Maximization in Diffusion Filtering. *Proc. of Int. Conf. on Artificial Neural Networks (ICANN 2005)*, Warsaw, Poland, September 2005. Lecture Notes in Computer Science 3697. Berlin 2005, Springer, pp. 247-252.

17. Girdziusas, R.; Laaksonen, J. Optimal Ratio of Lamé Moduli with Application to Motion of Jupiter Storms. *Proc. of 14th Scandinavian Conference on Image Analysis*, Joensuu, Finland, June 19-22, 2005. Springer-Verlag, pp. 1096-1106.

18. Girdziusas, R.; Laaksonen, J. Optimal Stopping and Constraints for Diffusion Models of Signals with Discontinuities. *Proc. of 16th European Conference on Machine Learning*, Porto, Portugal, October 3-7, 2005. pp. 576-583.

19. Girdziusas, R.; Laaksonen, J. Use of Input Deformations with Brownian Motion Filters for Discontinuous Regression. *Proc. of 3rd Int. Conf. on Advances in Pattern Recognition*, Bath, U.K., August 22-25, 2005. Springer-Verlag, pp. 219-228.

20. Hansen, J.H.L.; Huang, R.; Zhou, B.; Seadle, M.; Deller Jr., J.R.; Gurijala, A.R.; Kurimo, M.; Angkititrakul, P. SpeechFind: Advances in Spoken Document Retrieval for a National Gallery of the Spoken Word. *IEEE Transactions on Speech and Audio Processing*, 2005. Vol. 13, No. 5, pp. 712-730.

21. Harva, M.; Kabán, A. A Variational Bayesian Method for Rectified Factor Analysis. *Proc. of Int. Joint Conf. on Neural Networks (IJCNN'05)*, Montreal, Canada, July 31 - August 4, 2005. pp. 185-190.

22. Harva, M.; Raiko, T.; Honkela, A.; Valpola, H.; Karhunen, J. Bayes Blocks: An Implementation of the Variational Bayesian Building Blocks Framework. *Proc. of 21st Conference on Uncertainty in Artificial Intelligence*, Edinburgh, Scotland, July 26-29, 2005. pp. 259-266.

23. Hirsimäki, T.; Creutz, M.; Siivola, V.; Kurimo, M. Morphologically Motivated Language Models in Speech Recognition. *Proc. of International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR05)*, Espoo, Finland, June 15-17, 2005. pp. 121-126.

24. Honkela, A.; Valpola, H. Unsupervised Variational Bayesian Learning of Nonlinear Models. *Advances in Neural Information Processing Systems*, Vol. 17 (NIPS 2004, Vancouver & Whistler, BC, Canada, December 13-18, 2004). Cambridge, MA 2005, The MIT Press, pp. 593-600.

25. Honkela, A.; Östman, T.; Vigário, R. Empirical Evidence of the Linear Nature of Magnetoencephalograms. *Proc. of 13th European Symposium on Artificial Neural Networks (ESANN 2005)*, Bruges, Belgium, April 27-29, 2005. pp. 285-290.

26. Honkela, T. Von Foerster Meets Kohonen - Approaches to Artificial Intelligence, Cognitive Science and Information Systems Development. *Kybernetes*, 2005. Vol. 31, No. 1/2, pp. 40-53.

27. Honkela, T.; Hynnä, K.; Lagus, K.; Särelä, J. (eds.) Adaptive and Statistical Approaches in Conceptual Modeling. Espoo, Finland: Helsinki University of Technology, 2005. (Publications in Computer and Information Science Technical Report A75).

28. Honkela, T.; Hyvärinen, A.; Väyrynen, J. Emergence of Linguistic Features: Independent Component Analysis of Contexts. *Proc. of Ninth Neural Computation and Psychology Workshop: Modelling Language, Cognition and Action*, Plymouth, England, September 8-10, 2004. New Jersey 2005, World Scientific, pp. 129-138.

29. Honkela, T.; Könönen, V.; Pöllä, M.; Simula, O. (Eds.) *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05)*, Espoo, Finland, June 15-17, 2005. Espoo, Finland 2005. 174 p.

30. Honkela, T.; Saarikko, P. Research Nations. *Proceedings of the Summit of Micronations*, 2005. Helsinki, Finland 2005, MUU, pp. 29-33.

31. Hyvärinen, A.; Karhunen, J.; Oja, E. *Independent Component Analysis.* Japanese translation. Tokyo, Japan 2005, Tokyo Denki University Press.

32. Hyvönen, S.; Junninen, H.; Laakso, L.; Dal Maso, M.; Grönholm, T.; Bonn, B.; Keronen, P.; Aalto, P.; Hiltunen, V.; Pohja, T.; Launiainen, S.; Hari, P.; Mannila, H.; Kulmala, M. A Look at Aerosol Formation using Data Mining Techniques. *Atmos. Chem. Phys.*, 2005. Vol. 5, pp. 3345-3356.

33. Ilin, A.; Valpola, H. Frequency-Based Separation of Climate Signals. *Proc. of 9th European Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD 2005)*, Porto, Portugal, October 2005. pp. 519-526.

34. Ilin, A.; Valpola, H. On the Effect of the Form of the Posterior Approximation in Variational Learning of ICA Models. *Neural Processing Letters*, 2005. Vol. 22, No. 2, pp. 183-204.

35. Ilin, A.; Valpola, H.; Oja, E. Semiblind Source Separation of Climate Data Detects El Niño as the Component with the Highest Interannual Variability. *Proc. of Int. Joint Conf. on Neural Networks (IJCNN 2005)*, Montréal, Québec, Canada, August 2005. pp. 1722-1727.

36. Ji, Y.; Hao, J.; Reyhani, N.; Lendasse, A. Direct and Recursive Prediction of Time Series Using Mutual Information Selection. *Lecture Notes in Computer Science 3512,*

Computational Intelligence and Bioinspired Systems: 8th International Workshop on Artificial Neural Networks (IWANN 2005), Vilanova i la Geltrú, Barcelona, Spain, June 8-10, 2005. Berlin 2005, Springer, pp. 1010-1017.

37. Kaplan, A.Ya.; Fingelkurts, An.A.; Fingelkurts, Al.A.; Borisov, S.V.; Darkhovsky, B.S. Nonstationary Nature of the Brain Activity as Revealed by EEG/MEG: Methodological, Practical and Conceptual Challenges. *Signal Processing*, 2005. Vol. 85, pp. 2190-2212.

38. Kaski, S. From Learning Metrics Towards Dependency Exploration. *Proc. of 5th Workshop On Self-Organizing Maps (WSOM'05)*, Paris, France, September 5-8, 2005. pp. 307-314.

39. Kaski, S. Proactive Information Retrieval by Monitoring Eye Movements. *Abstracts of BrainIT, The Second International Conference on Brain-Inspired Information Technology*, Kyushu Institute of Technology, Kitakuyshu, Japan, October 7-9, 2005. p. 28.

40. Kaski, S.; Myllymäki, P.; Kojo, I. User Models from Implicit Feedback for Proactive Information Retrieval. *Proc. of Workshop 4 of the 10th International Conference on User Modeling; Machine Learning for User Modeling: Challenges*, Edinburgh, Scotland, July 24-25, 2005. pp. 25-26.

41. Kaski, S.; Nikkilä, J.; Savia, E.; Roos, C. Discriminative Clustering of Yeast Stress Response. In: Seiffert, U.; Jain, L.; Schweizer, P. (eds.), *Bioinformatics using Computational Intelligence Paradigms.* Berlin 2005, Springer, pp. 75-92.

42. Kaski, S.; Nikkilä, J.; Sinkkonen, J.; Lahti, L.; Knuuttila, J.; Roos, C. Associative Clustering for Exploring Dependencies between Functional Genomics Data Sets. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2005. Vol. 2, pp. 203-216.

43. Kaski, S.; Sinkkonen, J.; Klami, A. Discriminative Clustering. *Neurocomputing*, 2005. Vol. 69, pp. 18-41.

44. Kersting, K.; Raiko, T. 'Say EM' for Selecting Probabilistic Models for Logical Sequences. *Proc. of 21st Conference on Uncertainty in Artificial Intelligence (UAI 2005)*, Edinburgh, Scotland, July 26-29, 2005. pp. 300-307.

45. Kimmelma, O.; Parviainen, J.; Välimäki; V. Musiikkiäänitteiden digitaalinen vanhentaminen. *Suomen Musiikintutkijoiden 9. Valtakunnallinen Symposium*, Jyväskylä, Finland, March 17-19, 2005. pp. 55-58.

46. Klami, A.; Kaski, S. Non-Parametric Dependent Components. *Proc. of IEEE International Conference on Acoustis, Speech, and Signal Processing (ICASSP'05)*, Philadelphia, PA, USA, March 18-23, 2005. pp. V-209-V-212.

47. Kohonen, O.; Katajamäki, S.; Honkela, T. In Search for Volta: Statistical Analysis of Word Patterns in Shakespeare's Sonnets. *Proc. of International Symposium on Adaptive Models of Knowledge, Language and Cognition (AMKLC'05)*, Espoo, Finland, June 15-17, 2005. pp. 44-47.

48. Kohonen, T. Pointwise Organizing Projections. *Proc. of 5th Workshop On Self-Organizing Maps (WSOM'05)*, Paris, France, September 5-8, 2005. (CD-ROM)

49. Koldovský, Z.; Tichavský, P.; Oja, E. Cramér-Rao Lower Bound for Linear Independent Component Analysis. *Proc. of Int. Conf. on Acoustics, Speech, and Signal Proc. (ICASSP'05)*, Philadelphia, USA, March 20-23, 2005. Vol. III, pp. 581-584.

50. Koskela, M.; Laaksonen, J. Semantic Annotation of Image Groups with Self-Organizing Maps. *Proc. of 4th International Conference on Image and Video Retrieval (CIVR 2005)*, Singapore, July 2005. pp. 518-527.

51. Koskela, M.; Laaksonen, J.; Sjöberg, M.; Muurinen, H. PicSOM Experiments in TRECVID 2005. *Proc. of TRECVID 2005 Workshop*, Gaithersburg, MD, USA, November 2005. pp. 262-270.

52. Kurimo, M.; Turunen, V. Retrieving Speech Correctly Despite the Recognition Errors. *Proc. of 2nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms*, Edinburgh, UK, July 11-13, 2005. (http://groups.inf.ed.ac.uk/mlmi05/ ; 2-page abstract available online)

53. Kurimo, M.; Turunen, V. To Recover from Speech Recognition Errors in Spoken Document Retrieval. *Proc. of 9th European Conference on Speech Communication and Technology, Interspeech 2005*, Lisbon, Portugal, September 4-8, 2005. pp. 605-608.

54. Kurimo, M.; Turunen, V.; Ekman, I. Speech Transcription and Spoken Document Retrieval in Finnish. In: *Machine Learning for Multimodal Interaction, Revised Selected Papers of the MLMI 2004 Workshop*, Lecture Notes in Computer Science, Vol. 3361. Berlin 2005, Springer, pp. 253-262.

55. Könönen, V. Gradient Descent for Symmetric and Asymmetric Multiagent Reinforcement Learning. *Web Intelligence and Agent Systems: An International Journal (WIAS)*, 2005. Vol. 3, pp. 17-30.

56. Könönen, V. Hierarchical Multiagent Reinforcement Learning in Markov Games. *Proc. of International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR05)*, Espoo, Finland, June 15-17, 2005. pp. 71-77.

57. Laaksonen, J.; Viitaniemi, V.; Koskela, M. Application of Self-Organizing Maps and Automatic Image Segmentation to 101 Object Categories Database. *Proc. of Fourth International Workshop on Content-Based Multimedia Indexing (CBMI 2005)*, Riga, Latvia, June 2005.

58. Laaksonen, J.; Viitaniemi, V.; Koskela, M. Emergence of Semantic Concepts in Visual Databases. *Proc. of International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR 05)*, Espoo, Finland, June 2005. pp. 127134.

59. Lagus, K. Miten hermoverkkomallit selittävät kielen oppimista. Kielen ja kognition suhde, *Puheen ja kielen tutkimuksen yhdistyksen julkaisuja 37*, 2005 (Puheen ja kielen tutkimuksen yhdistyksen päivät, 17.-18.3. 2005).

60. Lagus, K.; Alhoniemi, E.; Seppä, J.; Honkela, A.; Wagner, P. Independent Variable Group Analysis in Learning Compact Representations for Data. *Proc. of International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05)*, Helsinki, Finland, June 15-17, 2005. pp. 49-56.

61. Lagus, K.; Creutz, M.; Virpioja, S. Latent Linguistic Codes for Morphemes using Independent Component Analysis. *Proc. of Ninth Neural Computation and Psychology Workshop: Modelling Language, Cognition and Action*, Plymouth, England, September 8-10, 2004. New Jersey etc. 2005, World Scientific.

62. Laiho, J.; Raivio, K.; Lehtimäki, P.; Hätönen, K.; Simula, O. Advanced Analysis Methods for 3G Cellular Networks. *IEEE Transactions on Wireless Communications*, 2005. Vol. 4, No. 3, pp. 930-942.

63. Lehtimäki, P.; Raivio, K. A Knowledge-Based Model for Analyzing GSM Network Performance. *Proc. of 6th International Symposium on Intelligent Data Analysis (IDA)*, Madrid, Spain, September 8-10, 2005. pp. 204-215.

64. Lehtimäki, P.; Raivio, K. A SOM Based Approach for Visualization of GSM Network Performance Data. *Proc. of 18th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems (IEA/AIE)*, Bari, Italy, June 22-25, 2005. pp. 588-598.

65. Lendasse, A.; François, D.; Wertz, V. Verleysen, M. Nonparametric Noise Estimation to Build Nonlinear Model in Chemometry. *Proc. of Chimiométrie 2005*, Lille, France, Nov. 30 - Dec. 1, 2005. pp. 44-47.

66. Lendasse, A.; Francois, D.; Wertz, V.; Verleysen, M. Nonlinear Time Series Prediction by Weighted Vector Quantization. *Future Generation Computer Systems*, 2005. Vol. 21, No. 7, pp. 1056-1067.

67. Lendasse, A.; Ji, Y.; Reyhani, N.; Verleysen, M. LS-SVM Hyperparameter Selection with a Nonparametric Noise Estimator. *Artificial Neural Networks: Formal Models and Their Applications, Lecture Notes in Computer Science 3697, International Conference on Artificial Neural Networks (ICANN'05)*, Warsaw, Poland, September 11-15, 2005. Berlin 2005, Springer, pp. 625-630.

68. Lendasse, A.; Simon, G.; Wertz, V.; Verleysen, M. Fast Bootstrap Methodology for Model Selection. *Neurocomputing*, 2005. Vol. 64, pp. 161-181.

69. Luyssaert, S.; Sulkava, M.; Raitio, H.; Hollmén, J. Are N and S Deposition Altering the Chemical Composition of Norway Spruce and Scots Pine Needles in Finland? *Environmental Pollution*, 2005. Vol. 138, No. 1, pp. 5-17.

70. Molinier, M.; Laaksonen, J.; Ahola, J.; Häme, T. Self-Organizing Map Application for Retrieval of Man-Made Structures in Remote Sensing Data. *Proc. of ESA-EUSC 2005: Image Information Mining Theory and Application to Earth Observation*, Frascat, Italy, October 2005.

71. Nikkilä, J.; Roos, C.; Kaski, S. Integration of Transcription Factor Binding and Gene Expression by Associative Custering. *Proc. of Symposium of Knowledge Representation in Bioinformatics (KRBIO05)*, Espoo, Finland, June 15-17, 2005. pp. 22-29.

72. Nikkilä, J.; Roos, C.; Savia, E.; Kaski, S. Explorative Modeling of Yeast Stress Response and its Regulation with gCCA and Associative Clustering. *International Journal of Neural Systems*, 2005. Vol. 15, No. 4, pp. 237-246.

73. Nowé, A.; Honkela, T.; Könönen, V.; Verbeeck, K. (eds.) *Proceedings of the Workshop W9 on Reinforcement Learning in Nonstationary Environments*, Porto, Portugal, 2005 (in conjunction with the 16th ECML and 9th PKDD, Oct. 3-7, 2005). Portugal 2005. 81 p.

74. Oja, E. Finding Hidden Factors in Large Spatiotemporal Data Sets. *Proc. of 2005 International Conference on Neural Networks and Brain*, Beijing, China, Oct. 13 - 15, 2005.

75. Oja, M.; Sperber, G. O.; Blomberg, J.; Kaski, S. Self-Organizing Map-Based Discovery and Visualization of Human Endogeneous Retroviral Sequence Groups. *International Journal of Neural Systems*, 2005. Vol. 15, No. 3, pp. 163-179.

76. Pakkanen, J. Examining the Behaviour of the Evolving Tree . *Proc. of 5th Workshop on Self-Organizing Maps (WSOM 05)*, Paris, France, September 5-8, 2005. pp. 163-170.

77. Pakkanen, J.; Iivarinen, J. Analyzing Large Image Databases with the Evolving Tree. *Proc. of Third International Conference on Advances in Pattern Recognition*, LNCS 3686. Bath, UK, August 22-25, 2005. Part I, pp. 192-198.

78. Pakkanen, J.; Iivarinen, J.; Oja, E. The Evolving Tree, a Hierarchical Tool for Unsupervised Data Analysis. *Proc. of International Joint Conference on Neural Networks*, Québec, Canada, July 31-August 4, 2005. pp. 1395-1399.

79. Peltonen, J.; Kaski, S. Discriminative Components of Data. *IEEE Transactions on Neural Networks*, 2005. Vol. 16, pp. 68-83.

80. Puolamäki, K.; Salojärvi, J.; Savia, E.; Simola, J.; Kaski, S. Combining Eye Movements and Collaborative Filtering for Proactive Information Retrieval. *SIGIR 2005: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Salvador, Brasil, August 15-19, 2005. New York, USA 2005, ACM Press, pp. 146-153.

81. Pylkkönen, J. An Efficient One-pass Decoder for Finnish Large Vocabulary Continuous Speech Recognition. *Proc. of 2nd Baltic Conference on Human Language Technologies (HLT'2005)*, Tallinn, Estonia, April 4-5, 2005. pp. 167-172.

82. Pylkkönen, J. New Pruning Criteria for Efficient Decoding. *Proc. of 9th European Conference on Speech Communication and Technology, Interspeech 2005*, Lisboa, Portugal, September 4-8, 2005. pp. 581-584.

83. Pöllä, M.; Lindh-Knuutila, T.; Honkela, T. Self-Refreshing SOM as a Semantic Memory Model. *Proc. of International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05)*, Espoo, Finland, June 15-17, 2005. pp. 171-174.

84. Raiko, T. Nonlinear Relational Markov Networks with an Application to the Game of Go. *Proc. of 15th International Conference on Artificial Neural Networks (ICANN 2005)*, Warsaw, Poland, September 11-15, 2005. Lecture Notes in Computer Science 3697. Berlin 2005, Springer, pp. 989-996.

85. Raiko, T.; Tornio, M. Learning Nonlinear State-Space Models for Control. *Proc. of International Joint Conference on Neural Networks (IJCNN 2005)*, Montreal, Canada, July 31-August 4, 2005. pp. 815-820.

86. Raju, K.; Huovinen, T.; and Ristaniemi, T. Blind Interference Cancellation Scheme for DS-CDMA Systems. *Proc. of IEEE International Symposium on Antennas and Propagation and USNC/URSI National Radio Science Meeting*, Washington DC, USA, July 3-8, 2005.

87. Rautkorpi, R.; Iivarinen, J. Contour Co-occurrence Matrix - A Novel Statistical Shape Descriptor. *Proc. of 13th International Conference on Image Analysis and Processing*, LNCS 3617, Cagliari, Italy, September 6-8, 2005. pp. 253-260.

88. Rautkorpi, R.; Iivarinen, J. Shape-Based Co-occurrence Matrices for Defect Classification. *Proc. of 14th Scandinavian Conference on Image Analysis*, LNCS 3540, Joensuu, Finland, June 19-22, 2005. Berlin 2005, Springer, pp. 588-597.

89. Reyhani, N.; Hao, J.; Ji, Y.; Lendasse, A. Mutual Information and Gamma Test for Input Selection. *Proc. of European Symposium on Artificial Neural Networks (ESANN 2005)*, Bruges, Belgium, April 27-29, 2005. pp. 503-504.

90. Roue, B.; Bas, P.; Chassery, J-M. Influence des Vecteurs Caractéristiques en Stéganalyse par Séparateurs à Vastes Marges. *Proc. of Gretsi 2005*, Louvain-la-Neuve, Belgique, Septembre 6-9, 2005.

91. Russell, A.; Honkela, T. Analysis of Interprofessional Collaboration in an Online Learning Environment using Self-Organizing Maps. *Proc. of International Symposium on Adaptive Models of Knowledge, Language and Cognition (AMKLC'05)*, Espoo, Finland, June 15-17, 2005. pp. 52-57.

92. Russell, A.; Honkela, T.; Lagus, K.; Pöllä, M. (eds.) *Proceedings of Symposium on Adaptive Models of Knowledge, Language and Cognition (AMKLC'05)*. Espoo, Finland, June 15-17, 2005. 61 p.

93. Salojärvi, J.; Puolamäki, K.; Kaski, S. Expectation Maximization Algorithms for Conditional Likelihoods. *Proc. of 22nd International Conference on Machine Learning (ICML 2005)*, Bonn, Germany, August 7-11, 2005. New York, USA 2005, ACM Press, pp. 753-760.

94. Salojärvi, J.; Puolamäki, K.; Kaski, S. Expectation Maximization Algorithms for Conditional Likelihoods. Espoo, Finland: Helsinki University of Technology, 2005. (Publications in Computer and Information Science Report A83).

95. Salojärvi, J.; Puolamäki, K.; Kaski, S. Implicit Relevance Feedback From Eye Movements. *Artificial Neural Networks: Biological Inspirations - ICANN 2005: 15th International Conference*, Warsaw, Poland, 11-15 September, 2005. Lecture Notes in Computer Science 3696. Berlin, Germany 2005, Springer-Verlag, pp. 513-518.

96. Salojärvi, J.; Puolamäki, K.; Kaski, S. On Discriminative Joint Density Modeling. *Machine Learning: ECML 2005, European Conference on Machine Learning*, Porto, Portugal, October 3-7, 2005. Lecture Notes in Artificial Intelligence 3270. Berlin, Germany 2005, Springer-Verlag, pp. 341-352.

97. Salojärvi, J.; Puolamäki, K.; Simola, J.; Kovanen, L.; Kojo, I; Kaski, S. Inferring Relevance from Eye Movements: Feature Extraction. Espoo, Finland: Helsinki University of Technology, 2005. 23 p. (Publications in Computer and Information Science Report A82).

98. Savia, E.; Puolamäki, K.; Sinkkonen, J.; Kaski, S. Two-Way Latent Grouping Model for User Preference Prediction. *Proc. of 21st Conference on Uncertainty in Artificial Intelligence (UAI 2005)*, Edinburgh, Scotland, July 26-29, 2005. pp. 518-525.

99. Siivola, V. Building Compact Language Models Incrementally. *Proc. of Second Baltic Conference on Human Language Technolgies*, Tallinn, Estonia, April 4-5, 2005. pp. 183-188.

100. Siivola, V.; Pellom, B. Growing an n-Gram Model. *Proc. of 9th European Conference on Speech Communication and Technology (Interspeech 2005)*, Lisboa, Portugal, September 4-8, 2005. pp. 1309-1312.

101. Similä, T. Self-Organizing Map Learning Nonlinearly Embedded Manifolds. *Information Visualization*, 2005. Vol. 4, No. 1, pp. 22-31.

102. Similä, T.; Laine, S. Visual Approach to Supervised Variable Selection by Self-Organizing Map. *International Journal of Neural Systems*, 2005. Vol. 15, No. 1-2, pp. 101-110.

103. Similä, T.; Tikka, J. Multiresponse Sparse Regression with Application to Multidimensional Scaling. *Proc. of 15th International Conference on Artificial Neural Networks (ICANN'05)*, Warsaw, Poland. September 11-15, 2005. Lecture Notes in Computer Science 3697. Berlin 2005, Springer, pp. 97-102.

104. Simon, G.; Lendasse, A.; Cottrell, C.; Fort, J.-C.; Verleysen, M. Time Series Forecasting: Obtaining Long Term Trends with Self-Organizing Maps. *Pattern Recognition Letters*, 2005. Vol. 26, No. 12, pp. 1795-1808.

105. Sinkkonen, J.; Kaski, S.; Nikkilä, L.; Lahti, L. Associative Clustering (AC): Technical Details. Espoo, Finland: Helsinki University of Technology, 2005. (Publications in Computer and Information Science Report A84).

106. Sirola, M.; Lampi, G.; Parviainen, J. SOM Based Decision Support in Failure Management. *Proc. of IEEE International Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS'2005)*, Sofia, Bulgaria, September 5-7, 2005. pp. 468-473.

107. Sirola, M.; Lampi, G.; Parviainen, J. SOM Based Decision Support in Failure Management. *International Journal of Computing*, 2005. Vol. 4, No. 3.

108. Sjöberg, M.; Laaksonen, J. Content-Based Retrieval of Web Pages and Other Hierarchical Objects with Self-Organizing Maps. *Artificial Neural Networks: Formal Models and Their Applications, International Conference on Artificial Neural Networks (ICANN'05)*, Warsaw, Poland, September 11-15, 2005. Lecture Notes in Computer Science 3697. Berlin 2005, Springer, pp. 841-846.

109. Sorjamaa, A.; Hao, J.; Lendasse, A. Mutual Information and k-Nearest Neighbors approximator for Time Series Predictions. *Artificial Neural Networks: Formal Models and Their Applications, International Conference on Artificial Neural Networks (ICANN'05)*, Warsaw, Poland, September 11-15, 2005. Lecture Notes in Computer Science 3697. Berlin 2005, Springer, pp. 553-558.

110. Sorjamaa, A.; Lendasse, A.; Verleysen, M. Pruned Lazy Learning Models for Time Series Prediction. *Proc. of European Symposium on Artificial Neural Networks (ESANN 2005)*, Bruges, Belgium, April 27-29, 2005. pp. 509-514.

111. Sorjamaa, A.; Reyhani, N.; Lendasse, A. Input and Structure Selection for *k*-NN Approximator. *Computational Intelligence and Bioinspired Systems: 8th International Workshop on Artificial Neural Networks (IWANN 2005)*, Vilanova i la Geltrú, Barcelona, Spain, June 8-10, 2005. Lecture Notes in Computer Science 3512. Berlin 2005, Springer, pp. 985-991.

112. Sulkava, M.; Rautio, P.; Hollmén, J. Combining Measurement Quality into Monitoring Trends in Foliar Nutrient Concentrations. *Artificial Neural Networks: Formal Models and Their Applications, International Conference on Artificial Neural Networks (ICANN'05)*, Warsaw, Poland, September 11-15, 2005. Lecture Notes in Computer Science 3697. Berlin 2005, Springer, pp. 761-767.

113. Särelä, J.; Valpola, H. Denoising Source Separation. *Journal of Machine Learning Research*, 2005. Vol. 6(Mar), pp. 233-272.

114. Tikka, J.; Hollmén, J.; Lendasse, A. Input Selection for Long-Term Prediction of Time Series. *Proc. of 8th International Work-Conference on Artificial Neural Networks (IWANN 2005)*, Barcelona, Spain, June 8-10, 2005. Berlin 2005, Springer-Verlag, pp. 1002-1009.

115. Venna, J.; Kaski, S. Local Multidimensional Scaling with Controlled Tradeoff Between Trustworthiness and Continuity. *Proc. of 5th Workshop On Self-Organizing Maps (WSOM'05)*, Paris, France, September 5-8, 2005. pp. 695-702.

116. Venna, J.; Kaski, S. Visualized Atlas of a Gene Expression Databank. *Proc. of Symposium of Knowledge Representation in Bioinformatics (KRBIO05)*, Espoo, Finland, June 15-17, 2005. pp. 30-36.

117. Viitaniemi, V.; Laaksonen, J. Keyword-Detection Approach to Automatic Image Annotation. *Proc. of 2nd European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies (EWIMT 2005)*, London, UK, November 2005. pp. 15-22.

118. Väyrynen, J.; Honkela, T. Comparison of Independent Component Analysis and Singular Value Decomposition in Word Context Analysis. *Proc. of International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05)*, Espoo, Finland, June 15-17, 2005. pp. 135-140.

119. Yang, Z.; Laaksonen, J. Approximated Cassification in Interactive Facial Image Retrieval. *Proc. of 14th Scandinavian Conference on Image Analysis (SCIA 2005)*, Joensuu, Finland, June 2005. Berlin, Germany 2005, Springer, pp. 770-779.

120. Yang, Z.; Laaksonen, J. Interactive Retrieval in Facial Image Database using Self-Organizing Maps. *Proc. of IAPR Conference on Machine Vision Applications (MVA 2005)*, Tsukuba Science City, Japan, May 2005. pp. 112-115.

121. Yang, Z.; Laaksonen, J. Partial Relevance in Interactive Facial Image Retrieval. *Proc. of 3rd International Conference on Advances in Pattern Recognition (ICAPR 2005)*, Bath, UK, August 2005. pp. 216-225.

122. Yuan, Z.; Oja, E. Projective Nonnegative Matrix Factorization for Image Compression and Feature Extraction. *Proc. of 14th Scandinavian Conference on Image Analysis (SCIA 2005)*, Joensuu, Finland, June 2005. Berlin, Germany 2005, Springer, pp. 333-342.

II From Data to Knowledge Research Unit Research Projects under the CIS Laboratory

Chapter 15

From Data to Knowledge Research Unit

Heikki Mannila, Jaakko Hollmén, Kai Puolamäki, Ella Bingham, Johan Himberg, Robert Gwadera, Hannes Heikinheimo, Antti Ukkonen, Jouni K. Seppänen, Nikolaj Tatti, Heli Hiisilä, Antti Rasinen, Mikko Korpela, Janne Toivola

15.1 Data mining at the Pattern Discovery group

The Pattern Discovery group in Otaniemi concentrates on combinations of pattern discovery and probabilistic modeling in data mining: pattern discovery aims at finding local phenomena, while modeling often aims at global analysis.

Pattern discovery techniques can be very efficient in finding frequently occurring patterns from large masses of data. Techniques for this task include both algorithmics in the traditional computer science sense and probabilistic methods.

The research topics include research on frequent itemsets, latent variable models and the problem of finding ordering from the data. The application areas include gene expression data, and paleontological and ecological data analysis.

The Pattern Discovery group in the Laboratory of Computer and Information Science is part of the From Data to Knowledge research unit, which is located in part in the Department of Computer Science at the University of Helsinki.

15.2 Discovering orderings

Hannes Heikinheimo, Heikki Mannila, Kai Puolamäki, Antti Ukkonen

In many applications, such as paleontology, medical genetics and ranking of user preferences, the 0-1 data has an underlying unknown order. In the case of paleontology, for example, we have a collection of fossil sites, which can be loosely regarded as a snapshot of the set of taxa that lived at a certain location at approximately same time. Sites and their taxa may be described as an 0-1 occurence matrix, like the one shown in figure 15.2, where the rows correspond to sites and the columns correspond to the taxa: one in entry (i, j) means that fossilized remnants of taxon j has been found at site i.

In paleontology seriation, or temporal ordering of fossil sites, is a central and difficult problem. The geological age determination is inaccurate, and some finds are classified inaccurately and many finds are missing. The underlying temporal order is partial; for example, two fossil sites may be from the same paleontological time period and hence this pair of sites has no preferred temporal order.

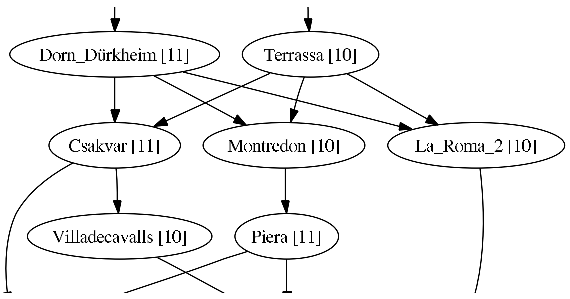


Figure 15.1: A subset of the discovered partial temporal order for 124 fossil sites, obtained by analyzing the occurence data on large late Cenzoic mammals.

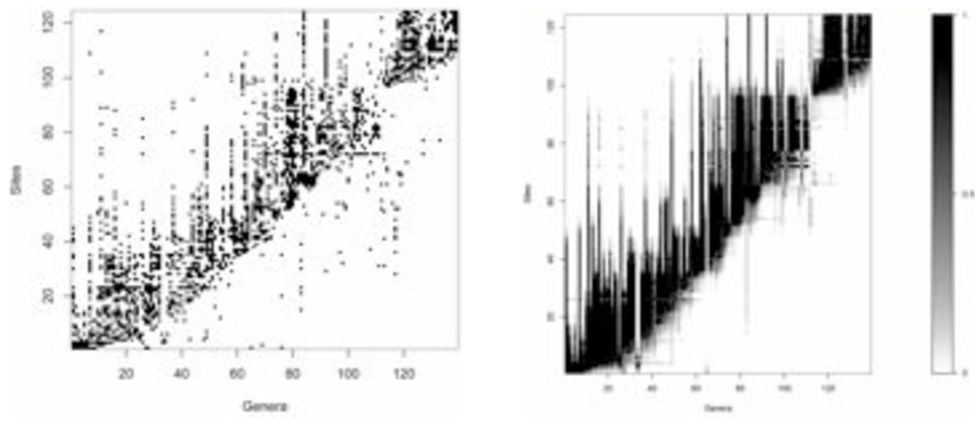


Figure 15.2: The figure on the left shows the original paleontological data in the preferred temporal order, given by the MCMC algorithm, black points denoting paleontological finds (ones in the occurence matrix). The figure on the right shows the probabilities that the genera were alive during the period of a find, given by our probabilistic model.

We have developed methods that can be used to find partial orders, based on the 0-1 occurence matrices. One approach is to study fragments of order, i.e., small subsets of items to be ordered, and find the preferred total order for the fragment by minimizing a score function. [3] In case of the paleontological data, a score function for a given ordering of sites can be for example the number of changes from 1 to 0 for a given genus. The fragments can then be used to construct a partial order. like one shown in figure 15.1.

We have also developed probabilistic models to solve the ordering problem. We have solved the models with Markov Chain Monte Carlo (MCMC) method, which is able to — unlike earlier methods used to analyze the seriation problem — actually produce a fine-grained temporal ordering. [2] Our method has also been used to identify false finds the fossil databases, and also genera with unusual ecological characteristic.

Another application of partial orders we are working on is the description and summarization of large sets of total orders (rankings). [3] Given several total orders of a set of items, it is possible to determine one or a few partial orders that describe the original total orders well.

In addition to the fossil data, our group is in the process of analyzing other large ecological datasets. [1] Via cluster analysis we have already gained some interesting and ecologically relevant results on spatial distributions of mammalian metacommunities in Europe. Furthermore, this study has risen theoretically interesting questions for further methodological studies on cluster analysis in the context of data with prevailing spatial relationships.

References

[1] Hannes Heikinheimo. Inferring taxonomic hierarchies from 0-1 data. Master's thesis, Helsinki University of Technology, 2005.

[2] Kai Puolamäki, Mikael Fortelius, and Heikki Mannila. Seriation in paleontological data using Markov Chain Monte Carlo methods. *PLoS Computational Biology*, 2(2):e6, February 2006. http://dx.doi.org/10.1371/journal.pcbi.0020006.

[3] Antti Ukkonen. Data mining techniques for discovering partial orders. Master's thesis, Helsinki University of Technology, 2004.

[4] Antti Ukkonen, Mikael Fortelius, and Heikki Mannila. Finding partial orders from unordered 0-1 data. In *Proceedings of the 11th ACM Conference on Knowledge Discovery and Data Mining (KDD)*, 2005.

15.3 Theoretical aspects of data mining

Jaakko Hollmén, Heikki Mannila, Jouni K. Seppänen, Nikolaj Tatti

Nowadays one often encounters high-dimensional data in practical applications. Thus methods and algorithms to analyse such data are highly needed. A typical example of such data is 0–1 data, i.e., the case when the data consists of vectors whose elements are 0 and 1. Although it seems that this kind of data is the simplest one, many applications include analysis of such data. For example, binary data can be generated from text documents such that each element of a binary vector represent some particular word, set to 1 if this word is present in a document and 0 otherwise. Different databases provide a large and important class of applications. For example, in market basket data each vector represent a transaction and the elements represent different products. More such examples can be obtained from course participation data or citation data. Binary data can be also obtained from genome data, e.g., single nucleotide polymorphisms (SNP) are a direct example of this.

Frequent itemsets are one of the best known concepts in 0–1 data mining: an itemset is frequent in a database if its items co-occur in sufficiently many records. Since the inception of frequent itemset mining as a solution to the association rule mining problem in 1996, several methods for finding all frequent itemsets have been proposed, and algorithms for this task continue to be a large research area within data mining. A question that we feel has not been satisfactorily addressed is that of using the itemsets: what do they tell us about the original data?

One way to answer this question is to use frequent itemsets for query approximation. Given a Boolean query ϕ over the attributes of the original data, how good approximations can one obtain using only the frequent itemsets? An answer that is in principle complete was given in [1]: the itemsets can be seen as the conditions of a linear program, whose objective function can be minimized or maximized to find the minimum and maximum of the Boolean query. This solution was shown by experiments to be useful in some cases, but it has the intrinsic problem that the size of the linear program is exponential in the number of variables. In fact, we show in [4] that such query problems are NP-complete. However, we show in [5] that under some assumptions we can drastically reduce the number of variables and thus ease the computational burden.

Another possibility is to use a combinatorial algorithm to approximate the query: for example, if the query is a disjunction of attributes,

$$\phi = A_1 \vee A_2 \vee \cdots \vee A_k,$$

its answer is an inclusion-exclusion sum

$$\sum_{j=1}^{k} f(A_j) - \sum_{1 \leq i < j \leq k} f(A_i A_j) + \sum_{1 \leq h < i < j \leq k} f(A_h A_i A_j) + \cdots \\ + (-1)^{k+1} f(A_1 A_2 \cdots A_k),$$

where f denotes the frequency of an itemset. Now if only the frequencies of frequent itemsets are filled in this sum, how far can it be from the correct result? The answer is twofold: in theory, the worst-case bound for the algorithm is very large, and a construction exists that shows the bound to be tight; but in practice, the approximations tend to be much closer to the correct answer than in the worst case. The theoretical part was addressed in [3], where the above approach was also extended to arbitrary Boolean formulas from the simple disjunctions. The practical results are as yet unpublished.

Finally, we comment in [2] on the recent idea of hypercube segmentation by Jon Kleinberg et al. Hypercube segmentation is one formalization of clustering 0–1 data. We show that the analysis by Kleinberg et al. of their algorithm is nearly tight, and if the approximation guarantee is to be significantly improved, the approach of selecting data vectors as cluster centers is not sufficient.

References

[1] Artur Bykowski, Jouni K. Seppänen, and Jaakko Hollmén. Model-independent bounding of the supports of boolean formulae in binary data. In Rosa Meo, Pier Luca Lanzi, and Mika Klemettinen, editors, *Database Support for Data Mining Applications: Discovering Knowledge with Inductive Queries*, volume 2682 of *Lecture Notes in Artificial Intelligence*, pages 234–249. Springer-Verlag, 2004.

[2] Jouni K. Seppänen. Upper bound for the approximation ratio of a class of hypercube segmentation algorithms. *Information Processing Letters*, 93(3):139–141, February 2005.

[3] Jouni K. Seppänen and Heikki Mannila. Boolean formulas and frequent sets. In Jean-François Boulicaut, Luc de Raedt, and Heikki Mannila, editors, *Constraint-based mining and inductive databases*, volume 3848 of *Lecture Notes in Artificial Intelligence*. Springer, 2005. To appear.

[4] Nikolaj Tatti. Computational complexity of queries based on itemsets. *Information Processing Letters*. In press.

[5] Nikolaj Tatti. Safe projections of binary data sets. *Acta Informatica*. In press. doi:10.1007/s00236-006-0009-9

15.4 Extending frequent itemsets: dense itemsets and tiles

Heikki Mannila, Jouni K. Seppänen

Another question related to frequent itemsets concerns extending their definition to relax the requirement of perfect co-occurrence: highly correlated items may form an interesting set, even if they never co-occur in a single record. The problem is to formalize this idea in a way that still admits efficient mining algorithms. Dense itemsets [2] are defined in a manner similar to frequent itemsets and can be found using a similar algorithm.

Another way to approach finding non-perfectly co-occurring items was defined in [1] and named "tiles". A spectral algorithm was used to rearrange the data matrix so that interesting sets of items become contiguous in both dimensions, and then these contiguous regions were found using a local search algorithm. This solution can also find non-perfectly anti-co-occurring items and hierarchical models where smaller tiles are used as exceptions to larger ones. An example of finding a hierarchical tile model is shown in Figure 15.3, where the leftmost pane shows the original data, the middle one shows the result of reordering, and the rightmost pane shows a model consisting of ten tiles.

An underlying theme connecting these topics is the interplay of two data mining objectives, local patterns and descriptive models. Frequent itemsets are an example of patterns: interesting phenomena occurring in some small part of the data. In contrast, a descriptive model tells us something interesting about the whole data. The query approximation problem is motivated by a desire to convert a frequent itemset collection into a model that can be used to answer queries about the data. Dense itemsets can be used to create a description of the data using e.g. a greedy algorithm on the dense itemsets. While frequent itemsets could be used similarly, the requirement of complete co-occurrence hinders the effort, and with dense itemsets the results are more interesting. Tile models are similarly descriptive models built from local patterns.

References

[1] Aristides Gionis, Heikki Mannila, and Jouni K. Seppänen. Geometric and combinatorial tiles in 0-1 data. In Jean-François Boulicaut, Floriana Esposito, Fosca Giannotti, and Dino Pedreschi, editors, *Knowledge Discovery in Databases: PKDD 2004*, volume 3202 of *Lecture Notes in Artificial Intelligence*, pages 173–184. Springer, 2004.

[2] Jouni K. Seppänen and Heikki Mannila. Dense itemsets. In Ronny Kohavi, Johannes Gehrke, William DuMouchel, and Joydeep Ghosh, editors, *Proceedings of the Tenth*

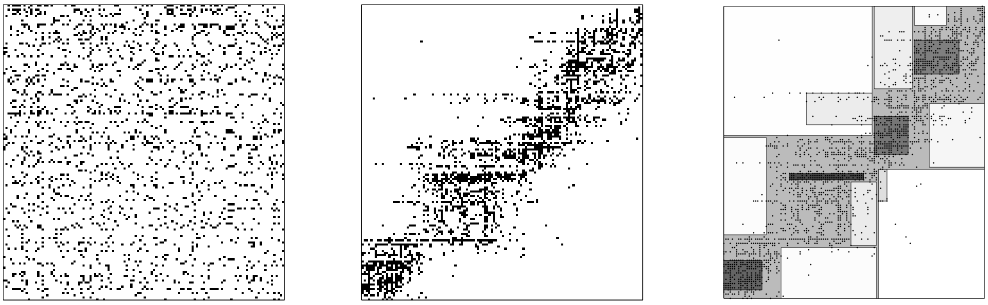


Figure 15.3: Example data first reordered, then hierarchically tiled

ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004), pages 683–688. ACM, 2004.

15.5 Basis segmentation

Ella Bingham, Heli Hiisilä, Heikki Mannila

In this project we have combined two techniques of multidimensional data analysis: segmentation of sequences, and dimensionality reduction. Both techniques reduce the complexity of representing the data. We use these techniques together, by finding the segment boundaries in a way that utilizes a low rank representation of the data.

We propose three different algorithms for the task, All of them consist of existing methods of segmentation (namely, optimal segmentation by dynamical programming) and dimensionality reduction (namely, principal component analysis which is optimal in the mean-squared-error sense).

- Algorithm SEG-PCA first segments the data, and then reduces the dimensionality of the data consisting of the segment means.
- Algorithm SEG-PCA-DP first segments the data, then reduces the dimensionality of the data consisting of the segment means, and then refines the segmentation by using the basis vectors of the dimensionality reduction.
- Algorithm PCA-SEG first decreases the dimensionality, and then segments the lower dimensional data.

As an application we might consider financial time series such as exchange rates of currencies. Each currency constitutes one dimension in the multidimensional data. Segmenting corresponds to splitting the time into different economical phases, and dimensionality reduction corresponds to finding dependencies between different currencies.

In a forthcoming paper [1] we have demonstrated our algorithms on exchange rate data, DNA sequences and meteorological time series.

References

[1] Ella Bingham, Aristides Gionis, Niina Haiminen, Heli Hiisilä, Heikki Mannila, Evimaria Terzi. Segmentation and dimensionality reduction. *2006 SIAM Conference on Data Mining, April 20-22, 2006, Bethesda, Maryland, USA.*

15.6 Data mining in bioinformatics

Ella Bingham, Heli Hiisilä, Johan Himberg, Jaakko Hollmén, Mikko Korpela, Heikki Mannila, Antti Rasinen, Jouni Seppänen, Janne Toivola

Bioinformatics is a new collaborative area of science that has risen out of the need to involve computer scientists in the analysis of data-intensive problems in biology and medicine. Data analysis plays an important role in gene expression studies, where high-dimensional, noisy microarray measurement matrices have to be processed to yield meaningful biological knowledge. This knowledge is needed in order to understand the functions of the genome on the whole and the role played by individual genes in particular diseases. This helps in developing diagnostic tools for early detection of diseases such as cancer [6], for instance. In addition to microarray measurements, auxiliary data set are helpful in reducing the uncertainty in the analysis. Auxiliary data sets include additional gene expression measurements using independent measurement platforms for validation, comparative genomic hybridization to measure gene copy number changes, tissue microarrays, and publicly available databases of gene expression and annotation databses, such as the gene ontology databases. Integration of different data sets is thus seen as an important aspect of data mining in bionformatics [2, 1].

In collaboration with researchers from the University of Helsinki and the Occupational Health Institute of Finland, we are involved in various projects involving cancer [4, 5, 6], also in analyzing cancer patients with work-related asbestos-exposure. Asbestos is a well known lung cancer causing mineral fiber.

Amplification profiling of human neoplasms In a recent study, we analyzed cancer-related gene amplification patterns collected from published literature. The data was collected at chromosome band-specific resolution from 838 published chromosomal comparative genomics hybridization studies for more than 4500 cases. We identified type-specific amplification profiles for each of the 73 cancer types (Fig. 15.4). Furthermore, relationships between the cancer types can be analyzed by a clustering solution relating the profiles with a similarity measure and performing a hierarchical clustering. In order to reveal generally interesting amplification patterns for cancer in general, we have identified amplification hot spots by means of independent component analysis. These genome-wide, sparse patterns of amplification offers an avenue for further exploration of genomic alterations of cancer.

Dependencies between transcription factor binding sites Gene expression of eucaryotes is regulated through transcription factors which are molecules able to attach to the binding sites in the DNA sequence. These binding sites are small pieces of DNA usually found upstream from the gene they regulate. As the binding sites play an important role in the gene expression, it is of interest to find out their characteristics.

In this project we look for dependencies and independencies between these binding sites using independent component analysis, non-negative matrix factorization, probabilistic latent semantic analysis and the method of frequent sets. The data used are human gene upstream regions and possible binding sites listed in a biological database. Also, data from the baker's yeast genome is analyzed. The results of the project are described in [3].

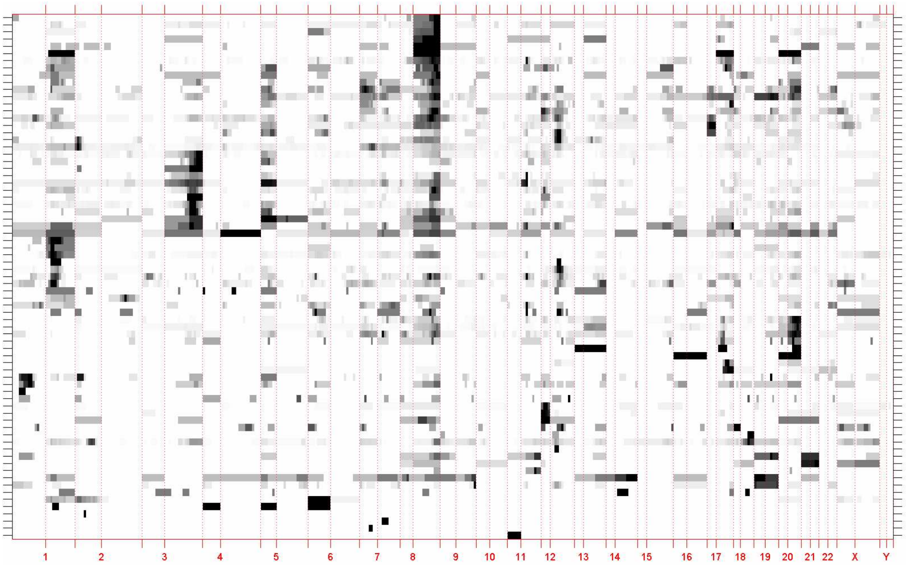


Figure 15.4: The cancer specific DNA amplification profiles are illustrated. The chromosomes are plotted on the horizontal axis one after the other in increasing order (starting with chromosome 1 on the left part of the figure). The profiles representing 73 cancer type profiles are plotted as rows.

References

[1] Catherine Bounsaythip, Erno Lindfors, Peddinti V. Gopalacharyulu, Jaakko Hollmén, and Matej Orešič. Network-based representation of biological data for enabling context-based mining. In *Proceedings of KRBIO'05, International Symposium of the Knowledge Representation in Bioinformatics*, pages 1–6, June 2005.

[2] Peddinti V. Gopalacharyulu, Erno Lindfors, Catherine Bounsaythip, Teemu Kivioja, Laxman Yetukuri, Jaakko Hollmén, and Matej Orešič. Data integration and visualization system for enabling conceptual biology. *Bioinformatics*, 21(Suppl.1):i177–i185, 2005.

[3] Heli Hiisilä and Ella Bingham. Dependencies between transcription factor binding sites: Comparison between ICA, NMF, PLSA and frequent sets. *Proceedings of the 4th IEEE International Conference on Data Mining, November 1-4, 2004, Brighton, UK*, pp. 114–121, 2004.

[4] Eeva Kettunen, Sisko Anttila, Jouni K. Sepp¡E4¿nen, Antti Karjalainen, Henrik Edgren, Irmeli Lindstr¡F6¿m, Reijo Salovaara, Anna-Maria Niss¡E9¿n, Jarmo Salo, Karin Mattson, Jaakko Hollmén, Sakari Knuutila, and Harriet Wikman. Differentially expressed genes in nonsmall cell lung cancer: expression profiling of cancer-related genes in squamous cell lung cancer. *Cancer Genetics and Cytogenetics*, 149(2):98–106, 2004.

[5] E. Kettunen, A.G. Nicholson, B. Nagy, J.K. Sepp¡E4¿nen, T. Ollikainen, G. Ladas, V. Kinnula, M. Dusmet, S. Nordling, J. Hollmén, D. Kamel, P. Goldstraw, and S. Knuutila. L1CAM, INP10, P-cadherin, tPA and ITGB4 over-expression in malignant pleural mesotheliomas revealed by combined use of cDNA and tissue microarray. *Carcinogenesis*, 26(1):17–25, 2005.

[6] Harriet Wikman, Jouni K. Sepp¡E4¿nen, Virinder K. Sarhadi, Eeva Kettunen, Kaisa Salmenkivi, Eeva Kuosma, Katri Vainio-Siukola, Balint Nagy, Antti Karjalainen, Thanos Sioris, Jarmo Salo, Jaakko Hollmén, Sakari Knuutila, and Sisko Anttila. Caveolins as tumor markers in lung cancer detected by combined use of cDNA and tissue microarrays. *Journal of Pathology*, 203:584–593, 2004.

Publications of the From Data to Knowledge Research Unit

Publications are in alphabetical order by the first author.

2004

1. Afrati, F.; Gionis, A.; Mannila, H. Approximating a Collection of Frequent Sets. *Proc. of 10th International Conference on Knowledge Discovery and Data Mining (KDD)*, Seattle, Washington, USA, Aug. 22-25, 2004.

2. Bykowski, A.; Seppänen, J.; Hollmén, J. Model-Independent Bounding of the Supports of Boolean Formulae in Binary Data. In: Meo, R., Lanzi, P. & Klemettinen, M. (eds.), *Database Support for Data Mining Applications: Discovering Knowledge with Inductive Queries.* Lecture Notes in Computer Science Vol. 2682. Heidelberg 2004, Springer-Verlag, pp. 234-249.

3. Geerts, F.; Mannila, H.; Terzi, E. Relational Link-Based Ranking. *Proc. of 30th International Conference on Very Large Data Bases (VLDB'04)*, Toronto, Canada, August 29 - September 3, 2004.

4. Gionis, A.; Mannila, H.; Seppänen, J. K. Geometric and Combinatorial Tiles in 0-1 Data. *Proc. of 8th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2004)*, Pisa, Italy, September 20-24, 2004. pp. 173-184.

5. Gionis, A.; Mannila, H.; Terzi, E. Clustered Segmentations. *Proc. of 3rd Workshop on Mining Temporal and Sequential Data (TDM)*, Seattle, Washington, USA, Aug. 22-25, 2004.

6. Hiisilä, H.; Bingham, E. Dependencies Between Tanscription Factor Binding Sites: Comparison Between ICA, NMF, PLSA and Frequent Sets. *Proc. of 4th IEEE International Conference on Data Mining*, Brighton, UK, November 1-4, 2004. pp. 114-121.

7. Himberg, J.; Hyvärinen, A.; Esposito, F. Validating the Independent Components of Neuroimaging Time-Series via Clustering and Visualization. *Neuroimage*, 2004. Vol. 22, No. 3, pp. 1214-1222.

8. Kaban, A.; Bingham, E.; Hirsimäki, T. Learning to Read Between the Lines: The Aspect Bernoulli Model. *Proc. of 4th SIAM International Conference on Data Mining*, Lake Buena Vista, Florida, USA, April 22-24, 2004. pp. 462-466.

9. Kettunen, E.; Anttila, S.; Seppänen, J. K.; Karjalainen, A.; Edgren, H.; Lindström, I.; Salovaara, R.; Nissén, A.-M.; Salo, J.; Mattson, K.; Hollmén, J.; Knuutila, S.; Wikman, H. Differentially Expressed Genes in Non-Small Cell Lung Cancer (NSCLC). Expression Profiling of Cancer-Related Genes in Squamous Cell Lung Cancer. *Cancer Genetics and Cytogenetics*, 2004. Vol. 149, No. 2, pp. 98-106.

10. Luyssaert, S.; Sulkava, M.; Raitio, H.; Hollmén, J. Evaluation of Forest Nutrition Based on Large-Scale Foliar Surveys: Are Nutrition Profiles the Way of the Future? *Journal of Environmental Monitoring*, 2004. Vol. 6, No. 2, pp. 160-167.

11. Mäntyjärvi, J.; Himberg, J.; Kangas, P.; Tuomela, U.; Huuskonen, P. Sensor Signal Data Set for Exploring Context Recognition of Mobile Devices. *Proc. of Workshop "Benchmarks and a database for context recognition"*, in conjuction with the 2nd Int. Conf. on Pervasive Computing (PERVASIVE 2004), Linz/Vienna, Austria, April 18-23, 2004. (Electronic publication, Swiss Federal Institute of Technology Zurich, Electronics laboratory)

12. Mäntyjärvi, J.; Nybergh, K.; Himberg, J.; Hjelt, K. Touch Detection System for Mobile Terminals. *Proc. of Mobile Human-Computer Interaction - MobileHCI 2004: 6th International Symposium*, Glasgow, UK, September 13-16, 2004. Heidelberg 2004, Springer-Verlag, pp. 331-336.

13. Patrikainen, A.; Mannila, H. Subspace Clustering of Binary Data - A Probabilistic Approach. *Proc. of SIAM Data Mining 2004, Workshop on Clustering High-Dimensional Data*, Lake Buena Vista, Florida, USA, April 22-24, 2004.

14. Puolamäki, K.,; Savia, E.; Sinkkonen, J.; Kaski, S. Two-Way Latent Grouping Model for User Preference Prediction. Espoo, Finland: Helsinki University of Technology, 2004. (Publications in Computer and Information Science, Report A80).

15. Salojärvi, J.; Puolamäki, K.; Kaski, S. Relevance Feedback from Eye Movements for Proactive Information Retrieval. *Proc. of Workshop on Processing Sensory Information for Proactive Systems (PSIPS 2004)*, Oulu, Finland, June 14-15, 2004. pp. 37-42.

16. Seppänen, J.; Mannila, H. Dense Itemsets. *Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004)*, Seattle, WA, USA, August 22-25, 2004. pp. 683-688.

17. Sulkava, M.; Tikka, J.; Hollmén, J. Sparse Regression for Analyzing the Development of Foliar Nutrient Concentrations in Coniferous Trees. *Proc. of Fourth International Workshop on Environmental Applications of Machine Learning (EAML 2004)*, Bled, Slovenia, September 2004. pp. 57-58.

18. Tikka, J.; Hollmén, J. Learning Linear Dependency Trees from Multivariate Time-series Data. *Proc. of Workshop on Temporal Data Mining: Algorithms, Theory and Applications*, in conjunction with The Fourth IEEE International Conference on Data Mining, Brighton, UK, November 2004.

19. Vesanto, J. & Hollmén, J. An Automated Report Generation Tool for the Data Understanding Phase. In: Abraham, A. , Jain, L. & van der Zwaag, B. J. (eds.), *Innovations in Intelligent Systems: Design, Management and Applications, Studies in Fuzziness and Soft Computing Vol. 140.* Heidelberg 2004, Springer (Physica) Verlag, chapter 5.

20. Wikman, H.; Seppänen, J. K.; Sarhadi, V. K.; Kettunen, E.; Salmenkivi, K.; Kuosma, E.; Vainio-Siukola, K.; Nagy, B.; Karjalainen, A.; Sioris, T.; Salo, J.; Hollmén, J.; Knuutila, S.; Anttila, S. Caveolins as Tumor Markers in Lung Cancer Detected by Combined Use of cDNA and Tissue Microarrays. *The American Journal of Pathology*, 2004. Vol. 203, pp. 584-593.

2005

1. Afrati, F.; Das, G.; Gionis, A.; Mannila, H.; Mielikäinen, T.; Tsaparas, P. Mining Chains of Relations. *Proc. of 5th International Conference on Data Mining, Houston (ICDM 2005)*, Texas, USA, November 27-30, 2005. pp. 553-556.

2. Boulicaut, J.-F.; de Raedt, L.; Mannila, H. (eds.) *Constraint-Based Mining and Inductive Databases*, Springer-Verlag LNCS Volume 3848. Berlin 2005, Springer-Verlag.

3. Bounsaythip, C.; Hollmén, J.; Kaski, S.; Oresic, M. (eds.) *Proceedings of KRBIO05, Symposium on Knowledge Representation in Bioinformatics*, Espoo, Finland, June 15-17, 2005. Espoo, Finland 2005, Helsinki University of Technology. 50 p.

4. Esposito, F.; Scarabino, T.; Hyvärinen, A.; Himberg, J.; Formisano, E.; Comani, S.; Tedeschi, G.; Goebel, R.; Seifritz, E.; Di Salle, F. Independent Component Analysis of fMRI Group Studies by Self-Organizing Clustering. *Neuroimage*, 2005. Vol. 25, No. 1, pp. 193-205.

5. Gionis, A.; Mannila, H.; Tsaparas, P. Clustering Aggregation. *Proc. of 21st International Conference on Data Engineering (ICDE 2005)*, Tokyo, Japan, April 5-8, 2005. pp. 341-352.

6. Hyvönen, S.; Junninen, H.; Laakso, L.; Dal Maso, M.; Grönholm, T.; Bonn, B.; Keronen, P.; Aalto, P.; Hiltunen, V.; Pohja, T.; Launiainen, S.; Hari, P.; Mannila, H.; Kulmala, M. A Look at Aerosol Formation using Data Mining Techniques. *Atmos. Chem. Phys.*, 2005. Vol. 5, pp. 3345-3356.

7. Kettunen, E.; Nicholson, A.G.; Nagy, B.; Wikman, H.; Seppänen, J.K.; Stjernvall, T.; Ollikainen, T.; Kinnula, V.; Nordling, S.; Hollmén, J.; Anttila, S.; Knuutila, S. L1CAM, INP10, P-cadherin, tPA and ITGB4 Over-Expression in Malignant Pleural Mesotheliomas Revealed by Combined Use of cDNA and Tissue Microarray. *Carcinogenesis*, 2005. Vol. 26, No. 1, pp. 17-25.

8. Luyssaert, S.; Sulkava, M.; Raitio, H.; Hollmén, J. Are N and S Deposition Altering the Chemical Composition of Norway Spruce and Scots Pine Needles in Finland? *Environmental Pollution*, 2005. Vol. 138, No. 1, pp. 5-17.

9. Mannila, H.; Salmenkivi, M. Piecewise Constant Modeling of Sequential Data Using Reversible Jump Markov Chain Monte Carlo. In: Wang, J.; Zaki, M.; Toivonen, H.; Shasha, D. (eds.), *Data Mining in Bioinformatics.* Berlin 2005, Springer, pp. 85-103.

10. Papadimitriou, S.; Gionis, A.; Tsaparas, P.; Vaisanen, R.A.; Mannila, H.; Faloutsos, C. Parameter-Free Spatial Data Mining Using MDL. *Proc. of 5th International Conference on Data Mining (ICDM 2005)*, Houston, Texas, USA, November 27-30, 2005. pp. 346-353.

11. Puolamäki, K.; Salojärvi, J.; Savia, E.; Simola, J.; Kaski, S. Combining Eye Movements and Collaborative Filtering for Proactive Information Retrieval. *SIGIR 2005: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Salvador, Brasil, August 15-19, 2005. New York, USA 2005, ACM Press, pp. 146-153.

12. Rastas, P.; Koivisto, M.; Mannila, H.; Ukkonen, E. A Hidden Markov Technique for Haplotype Reconstruction. *Proc. of Algorithms in Bioinformatics: 5th International Workshop (WABI 2005)*, Lecture Notes in Computer Science, 3692. Berlin 2005, Springer, pp. 140-151.

13. Salmenkivi, M.; Mannila, H. Using Markov Chain Monte Carlo and Dynamic Programming for Event Sequence Data. *Knowl. Inf. Syst.*, 2005. Vol. 7, No. 3, pp. 267-288.

14. Salojärvi, J.; Puolamäki, K.; Kaski, S. Expectation Maximization Algorithms for Conditional Likelihoods. *Proc. of 22nd International Conference on Machine Learning (ICML 2005)*, Bonn, Germany, August 7-11, 2005. New York, USA 2005, ACM Press, pp. 753-760.

15. Salojärvi, J.; Puolamäki, K.; Kaski, S. Expectation Maximization Algorithms for Conditional Likelihoods. Espoo, Finland: Helsinki University of Technology, 2005. (Publications in Computer and Information Science Report A83).

16. Salojärvi, J.; Puolamäki, K.; Kaski, S. Implicit Relevance Feedback From Eye Movements. *Artificial Neural Networks: Biological Inspirations - ICANN 2005: 15th International Conference*, Warsaw, Poland, 11-15 September, 2005. Lecture Notes in Computer Science 3696. Berlin, Germany 2005, Springer-Verlag, pp. 513-518.

17. Salojärvi, J.; Puolamäki, K.; Kaski, S. On Discriminative Joint Density Modeling. *Machine Learning: ECML 2005, European Conference on Machine Learning*, Porto, Portugal, October 3-7, 2005. Lecture Notes in Artificial Intelligence 3270. Berlin, Germany 2005, Springer-Verlag, pp. 341-352.

18. Salojärvi, J.; Puolamäki, K.; Simola, J.; Kovanen, L.; Kojo, I; Kaski, S. Inferring Relevance from Eye Movements: Feature Extraction. Espoo, Finland: Helsinki University of Technology, 2005. 23 p. (Publications in Computer and Information Science Report A82).

19. Savia, E.; Puolamäki, K.; Sinkkonen, J.; Kaski, S. Two-Way Latent Grouping Model for User Preference Prediction. *Proc. of 21st Conference on Uncertainty in Artificial Intelligence (UAI 2005)*, Edinburgh, Scotland, July 26-29, 2005. pp. 518-525.

20. Seppänen, J. Upper Bound for the Approximation Ratio of a Class of Hypercube Segmentation Algorithms. *Information Processing Letters*, 2005. Vol. 93, No. 3, pp. 139-141.

21. Seppänen, J.K.; Mannila, H. Boolean Formulas and Frequent Sets. In: Boulicaut, J.-C.; de Raedt, L.; Mannila, H. (eds.), *Constraint-Based Mining and Inductive Databases*, LNCS Volume 3848. Berlin 2005, Springer-Verlag, pp. 348-361.

22. Sulkava, M.; Rautio, P.; Hollmén, J. Combining Measurement Quality into Monitoring Trends in Foliar Nutrient Concentrations. *Artificial Neural Networks: Formal Models and Their Applications, International Conference on Artificial Neural*

Networks (ICANN'05), Warsaw, Poland, September 11-15, 2005. Lecture Notes in Computer Science 3697. Berlin 2005, Springer, pp. 761-767.

23. Tikka, J.; Hollmén, J.; Lendasse, A. Input Selection for Long-Term Prediction of Time Series. *Proc. of 8th International Work-Conference on Artificial Neural Networks (IWANN 2005)*, Barcelona, Spain, June 8-10, 2005. Berlin 2005, Springer-Verlag, pp. 1002-1009.

24. Ukkonen, A.; Fortelius, M.; Mannila, H. Finding Partial Orders from Unordered 0-1 Data. *Proc. of Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Chicago, Illinois, USA, August 21-24, 2005. pp. 285-293.