BIENNIAL REPORT

2002 – 2003

Laboratory of Computer and Information Science

Neural Networks Research Centre

Helsinki University of Technology

P.O. Box 5400

FI-02015 HUT, Finland

K. Puolamäki and L. Koivisto, editors

Otaniemi, February 2004

ISBN 951-22-6973-2
ISBN 951-22-6974-0 (electronic version)

Yliopistopaino
Helsinki 2004

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 the 1960's 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 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 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 also implies financial resources from the Academy, Tekes, and HUT, which are gratefully acknowledged.

The Neural Networks Research Centre operates 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 LCIS also participate in some research project of the NNRC. Professor Erkki Oja is presently the director of NNRC, with Professor Olli Simula the vice-director, and Professor Juha Karhunen participating in its research projects. In addition, 18 post-doctoral researchers, 34 graduate students, and a number of undergraduate students are working in the NNRC projects.

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 are financially separate and stem from different research traditions, there is an overlap in the research directions and projects between these two Centers of Excellence. This overlap has already produced fruitful joint research which is expected to increase in the future.

The present report covers the activities during the years 2002 and 2003. 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 status of the NNRC, up to the end of 2001, was thoroughly explained in the triennial reports 1994 - 1996 and 1997 - 1999 as well as the biennial report 2000 - 2001. 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 2002 - 2003, the laboratory produced 8 D.Sc. (Eng.) degrees, 2 Lic.Tech. degrees, and 33 M.Sc. (Eng.) degrees. The number of scientific publications appearing during the period was 229, of which 59 were journal papers. Compared to the previous two-year period 2000 - 2001, the number of doctoral degrees increased by 2, the number of publications by 35, and the number of journal papers by 21. 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. During the reporting period, Professor Erkki Oja was President of the European Neural Network Society. Also, some prices and honours, both national and international, were granted to members of our staff.

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

Personnel

Employees during 2002 – 2003

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.)
Heikki Mannila, D.Phil. Vice Director, From Data to Knowledge research unit
Jaakko Hollmén, D.Sc. (Tech.) (Acting Professor, H. Mannila's chair)
Mikko Kurimo, D.Sc. (Tech.) until July 31, 2003 (Acting Professor, E. Oja's chair)
Timo Honkela, D.Phil. from Aug. 1, 2003 (Acting Professor, E. Oja's chair)
Petteri Pajunen, D.Sc. (Tech.) until July 31, 2002 (Acting Professor, J. Karhunen's chair)

Post-doc researchers

Ella Bingham, D.Sc. (Tech.)
Timo Honkela, D.Phil.
Patrik Hoyer, D.Sc. (Tech.) (visiting abroad from Jan. 2003)
Jarmo Hurri
Aapo Hyvärinen, D.Phil. (until May 31, 2003)
Jukka Iivarinen, D.Sc. (Tech.)
Sirkka-Liisa Joutsiniemi, D.Med.Sc. (until May 31, 2003)
Samuel Kaski, D.Sc. (Tech.)
Markus Koskela, D.Sc. (Tech.)
Mikko Kurimo, D.Sc. (Tech.)
Maija-Liisa Laakso, D.Med.Sc. (until Dec. 31, 2002)
Jorma Laaksonen, D.Sc. (Tech.)
Krista Lagus, D.Sc. (Tech.)
Sampsa Laine, D.Sc. (Tech.)
Petteri Pajunen, D.Sc. (Tech.)
Kai Puolamäki, D.Phil.
Kimmo Raivio, D.Sc. (Tech.)
Janne Sinkkonen, D.Sc. (Tech.)
Miki Sirola, D.Sc. (Tech.), laboratory engineer
Panu Somervuo, D.Sc. (Tech.)
Harri Valpola, D.Sc. (Tech.) (visiting abroad from Sept. 1, 2003)
Ricardo Vigário, D.Sc. (Tech.)
Vuokko Vuori (until Dec. 31, 2002)

Post-graduate researchers

Matti Aksela
Esa Alhoniemi (until Sept. 30, 2002)
Mathias Creutz
Ramunas Girdziusas
Johan Himberg (absent March 2000 – Oct. 2002)
Teemu Hirsimäki
Antti Honkela
Alexander Ilin
Mika Inki
Heli Juntunen (from June 1, 2002)
Mikko Katajamaa
Arto Klami
Mikko Koivisto
Ville Könönen
Leo Lahti (from Feb. 17, 2003)
Pasi Lehtimäki
Janne Nikkilä
Merja Oja
Jussi Pakkanen
Jukka Parviainen
Anne Patrikainen
Jaakko Peltonen
Tapani Raiko
Karthikesh Raju
Mika Rummukainen (Oct. 9, 2002 – July 31, 2003)
Salla Ruosaari
Jarkko Salojärvi
Eeva Savia (from March 5, 2003)
Jouni Seppänen
Vesa Siivola
Mika Sulkava
Jaakko Särelä
Zhirong Yang (from Nov. 1, 2003)
Zhijian Yuan (from July 1, 2003)
Jarkko Venna
Sampo Viiperi
Ville Viitaniemi

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

Markus Harva
Mikko Heikelä
Kevin Hynnä
Jari-Pekka Juhala
Oskar Kohonen
Mikko Korpela
Kalle Korpiaho
Jukka Kuusisto
Golan Lampi

Leo Lundqvist
Teppo Marin
Teemu Murtola
Vesa Paatero
Janne Pylkkönen
Antti Rasinen
Rami Rautkorpi
Leah Russell
Juha Räisänen
Toni Saarela
Petri Saarikko
Antti Savolainen
Jan-Hendrik Schleimer
Timo Similä
Mats Sjöberg
Nikolai Tatti
Johanna Tikanmäki
Jarkko Tikka
Janne Toivola
Ville Turunen
Sami Virpioja
Jaakko Väyrynen
Paul Wagner
Juha Winter
Jarkko Ylipaavalniemi
Tomas Östman

Support staff

Teddy Grenman, maintenance assistant
Leila Koivisto, department secretary
Sakari Laitinen, maintenance assistant
Tarja Pihamaa, laboratory secretary
Teemu Pohjolainen, maintenance assistant
Markku Ranta, B.Eng., works engineer
Petteri Räisänen, maintenance assistant
Yujie Ye, maintenance assistant

Awards and activities

Prizes and scientific honours received by researchers of the unit

M.Sc. Antti Honkela:

- Master's Thesis Award on Mathematical and Computational Modeling, Center for Mathematical and Computational Modeling, University of Jyväskylä. 2002.

Dr. Samuel Kaski:

- IEEE Senior Member. 2002.

Dr. Jorma Laaksonen:

- IEEE Senior Member. 2002.

Professor Heikki Mannila:

- ACM SIGKDD Innovation Award, ACM SIGKDD. 2003.

Dr. Jorma Laaksonen, M.Sc. Markus Koskela, Professor Erkki Oja:

- Best Paper Award at the Workshop on Self-Organizing Maps, Japan. 2003.

M.Sc. Jaakko Peltonen:

- Master's Thesis Award, The Finnish Association of Graduate Engineers TEK. 2002.

Dr. Harri Valpola:

- Dissertation Award, The Finnish Pattern Recognition Society. 2002.

Important international positions of trust held by researchers of the unit

Academy Professor Erkki Oja:

- Editorial Board Member:
  Neural Computation (USA)
  Pattern Recognition Letters (The Netherlands) –2002
  International Journal of Pattern Recognition and and Artificial Intelligence (Singapore)
  International Journal of Computer Research (USA)
  Journal of Natural Computing (The Netherlands).
- Editor, *Artificial Neural Networks and Neural Information Processing*, Lecture Notes in Computer Science 2714, Springer, Berlin, 2003.
- European Neural Network Society (ENNS), The Netherlands. President.
- Evaluator of EU projects, Belgium, 2003.
- International Neural Network Society, USA. Governing Board Member.
- IEEE - Neural Networks Chapter, Finland. Chairman.
- Evaluator in filling the academic chairs of professors:
  computational neuroscience, Universität Osnabruck, Germany, 2002
  computer science, University of Strathclyde, U.K., 2002
  electrical and computer engineering, University of Houston, USA, 2002
  electrical and computer engineering, University of Minnesota, USA, 2002
  associate research scientist (professor), University of California San Diego, USA, 2003.
- Opponent at the doctoral dissertation of Marc Van Hulle, Danish Technical University, Denmark, 2002.
- Opponent at the doctoral dissertation of Martin Kermit, University of Oslo, Norway, 2003.
- Plenary talk "Independent Component Analysis: introduction." European meeting on ICA, Vietri sul Mare, Italy, Feb. 21- 23, 2002.
- Tutorial talk "The Self-Organizing Map: basic theory." Natural Computing Days, Newcastle upon Tyne, England, March 4-5, 2002.
- Tutorial talk "Independent Component Analysis: basic theory." Natural Computing Days, Newcastle upon Tyne, England, March 4-5, 2002.
- Plenary talk "Independent Component Analysis: recent adcvances." Int. Conf. on Cognitive and Neural Systems (ICCNS), Boston, USA, May 28-31, 2002.
- Plenary talk "Independent Component Analysis: fundamentals and recent advances." Euro-International Symposium on Computational Intelligence (E-ISCI), Kosice, Slovakia, June 17-19, 2002.
- Plenary talk "Independent Component Analysis: fundamentals and recent advances." 26th Annual Conf. of the Gesellschaft fur Klassifikation, Mannheim, Germany, July 22-24, 2002.

- Plenary talk "Finding hidden factors using Independent Component Analysis." European Conference on Machine Learning (ECML), Helsinki, Finland, August 20-23, 2002.
- Plenary talk "Independent Component Analysis." Hybrid Intelligent Systems (HIS'03), Santiago de Chile, Chile, Dec. 1-4, 2002.
- Tutorial talk "Data and image mining with SOM." 4th Workshop on Self-Organizing Maps (WSOM), Hibikino, Japan, Sept. 11-14, 2003.
- Program Chairman, 2nd Euro-International Symposium on Computational Intelligence, Kosice, Slovakia, June 16-19, 2002.
- Program Chairman and Session Chairman:
  13th Int. Conf. on Artificial Neural Networks, ICANN'03, Istanbul, Turkey, June 26-29, 2003
  Workshop on Self-Organizing Maps, WSOM'03, Hibikino, Japan, Sept. 11-14, 2003.
- Session Chairman and Program Committee Member:
  4th Int. Symposium on Independent Component Analysis and Blind Source Separation, Nara, Japan, April 1-4, 2003
  Int. Joint Conf. on Neural Networks, Portland, USA, July 21-24, 2003.
- Session Chairman:
  World Congress on Computational Intelligence, Honolulu, USA, May 11-17, 2002
  Between Data Science and Everyday Web Practice, Mannheim, Germany, July 22-24, 2002
  Workshop on New Directions for Signal Processing in the 21st Century, Lake Louise, Canada, October 5-10, 2003.
- Program Committee Member:
  Int. Conf. on Artificial Neural Networks, Madrid, Spain, August 27-30, 2002
  Int. Conf. on Neural Information Processing, Singapore, November 18-22, 2002.

Academician Teuvo Kohonen:

- Evaluator in filling the academic chair of Associate Professor, Drexel University, USA, 2002.
- Plenary talk "Self-organized maps of sensory events." Nobel Symposium, Stockholm, Sweden, August 25-27, 2002.
- Plenary talk "A computational model of visual attention." Int. Conf. on Artificial Neural Networks and Genetic Algorithms, ICANNGA'03, Roanne, France, April 23-25, 2003.
- Plenary talk "A computational model of visual attention." 13th Int. Conf. on Artificial Neural Networks, ICANN'03, Istanbul, Turkey, June 26-29, 2003.
- Invited talk "A computational model of visual attention." Int. Joint Conf. on Neural Networks, IJCNN'03, Portland, Oregon, USA, July 20-24, 2003.
- Tutorial talk "The SOM: How it was invented, what is its connection to the brain, and how can it be generalized." Workshop on Self-Organizing Maps, WSOM'03, Hibikino, Japan, Sept. 11-14, 2003.

Professor Juha Karhunen:

- Editorial Board Member:
  Neurocomputing (The Netherlands)
  Neural Processing Letters (The Netherlands)
- Invited poster presentation "Constructing graphical models for Bayesian ensemble learning from simple building blocks." Workshop on Learning, Snowbird, Utah, USA, April 2-5, 2002.
- Invited talk "An ensemble learning approach to nonlinear dynamic blind source separation using state-space models." The 2002 Int. Joint Conf. on Neural Networks (IJCNN2002), Honolulu, Hawaii, May 2002 (in special invited session on advances in independent component analysis, presented by A. Honkela).
- Invited talk "Advances in nonlinear blind source separation." The 4th Int. Workshop on Independent Component Analysis and Blind Source Separation (ICA2003), Nara, Japan, April 1-4, 2003 (in a special session on nonlinear independent component analysis and blind signal separation, presented by Prof. C. Jutten).
- Invited talk "Bayesian modeling of variance sources." European Summer School on ICA, Berlin, Germany, June 16-17, 2003 (presented by T. Raiko).
- Session Chairman and Program Committee Member:
  4th Int. Symp. on Independent Component Analysis and Blind Signal Separation (ICA2003), Nara, Japan, April 1-3, 2003.
- Program Committee Member:
  IEEE ISCAS2002, Phoenix, Arizona, USA, May, 26-29, 2002
  IEEE IJCNN2002 (WCCI2002), Honolulu, Hawaii, USA, May, 12-17, 2002
  ESANN2002, Bruges, Belgium, April 24-26, 2002
  IEEE ISCAS2003, Bangkok, Thailand, May 26-29, 2003

Professor Olli Simula:

- European Neural Network Society (ENNS), Great Britain. Governing Board Member.
- IEEE - Computer Chapter (C-16), Finland. Chairman.
- Plenary talk "SOM based analysis of industrial processes and telecommunication systems." Workshop on Self-Organizing Maps (WSOM'03), Hibikino, Japan, Sept. 11-14, 2003.
- Session Chairman and Program Committee Member:
  Int. Conf. on Artificial Neural Networks, ICANN 2002, Madrid, Spain, Aug. 28-30, 2002
  Workshop on Self-Organizing Maps, WSOM'03, Hibikino, Japan, Sept. 11-14, 2003.
- Program Committee Member, 13th Int. Conf. on Artificial Neural Networks, ICANN'03, Istanbul, Turkey, June 26-29, 2003.

Professor Heikki Mannila:

- Editor-in-Chief, Data Mining and Knowledge Discovery (USA).

- Associate Editor, ACM Transactions on Internet Technology (USA).
- Action Editor, Journal of Machine Learning Research (USA).
- Area Editor, IEEE Transactions on Knowledge and Data Engineering (USA).
- Member of PKDD Steering Committee.
- Member of ACM SIGKDD Curriculum Committee 2003–.
- Member of Technical Advisory Board, Verity Inc.
- Plenary talk, "Global structure from sequences." IEEE Int. Conf. on Data Mining (ICDM 2003), Melbourne, Fl., USA, 2003.
- Co-Chairman:
  Sixth European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD 2002), Helsinki, Finland
  Machine Learning: ECML 2002 - 12th European Conference on Machine Learning, Helsinki, Finland.
- Program Co-Chairman, Second SIAM Int. Conf. on Data Mining 2002, San Francisco, CA, USA.
- Program Committee Member:
  IEEE Int. Conf. on Data Mining (ICDM 2002), Macbashi, Japan
  29th Int. Colloquium on Automata, Languages and Programming - ICALP 2002, Malaga, Spain
  Discovery Science 2002, Lübeck, Germany
  Eighth SIGKDD Conference on Data Mining and Knowledge Discovery (KDD'02), Edmonton, Canada
  Int. Conf. on Database Theory (ICDT 2003), Siena, Italy
  ACM Symposium on Management of Data (SIGMOD 2003), San Diego, CA, USA
  IEEE Int. Conf. on Data Mining (ICDM 2003), Melbourne, Florida, USA
  Int. Conf. on Machine Learning (ICML 2003), Washington, D.C., USA
  15th Int. Conf. on Scientific and Statistical Database Management (SSDBM 03), Cambridge, MA, USA.

Professor Jaakko Hollmén:

- Member of the Management Board in the Knowledge Discovery Network of Excellence (KDNet) supported by EU Project No. IST-2001-33086.
- Member of the Management Board of the research project "Smart information system for waste treatment (iWaste)" in the STREAMS Technology project funded by TEKES.
- Invited talk "Data analysis of 0-1 data by combining frequent sets and mixture models." Teoriapäevad Pedasel (Theoretical computer science days), Pedase, Estonia, Oct. 3-5, 2003.
- Invited talk "Bioinformatics – something for the computer scientists?" Nordic University Computer Club Conference (NUCCC 2003), Espoo, Finland, March 28-30, 2003.

- Invited talk "Analysis of microarray data: current research and future challenges." Microarray data analysis developers' days. CSC – Scientific Computing, Espoo, Finland, May 26-27, 2003.
- Invited talk "Analysis of microarray data." Graduate school on microarrays, Helsinki Biomedical Graduate School, Helsinki, Finland, May 19-21, 2003.
- Invited talk "Beyond clustering: case studies and possibilities in gene expression data analysis." Microarray Bioinformatics Seminar, The Joint Bioinformatics Lab. of Turku Centre for Computer Science and Turku Centre for Biotechnology, Turku, Finland, May 6-7, 2003.
- Program Committee Member, Workshop on Self-Organizing Maps, WSOM'03, Hibikino, Japan, Sept. 11-14, 2003.

Professor Timo Honkela

- International Federation on Information Processing (IFIP), TC12 (Artificial Intelligence), Representative of Finland.
- Expert for Research Programme on Proactive Computing, Academy of Finland and French Ministry of Research, member of evaluation panel, 2002.
- Project Evaluator for the Sixth Framework Programme of EU Commission, 2003.
- Invited talk "Visual data and text mining for medical education, research and practice." Karolinska University, Stockholm, Sweden, 22nd of April, 2003.
- Invited talk "Emergence of implicit and explicit categories: cognitive models based on self-organizing maps and independent component analysis." Int. Conf. on Learning and Concept Formation, Lund University, Sweden, 14th of June, 2003.

Dr. Aapo Hyvärinen:

- Editor-in-Charge, International Journal of Neural Systems (Singapore).
- Plenary talk "Extensions of ICA as models of natural images and visual processing." Int. Symposium on Independent Component Analysis and Blind Source Separation, Nara, Japan, April 1-4, 2003.
- Invited talk "Activity bubbles and natural image sequences." Int. Conf. on Knowledge-Based Intelligent Information and Engineering Systems (KES2002), Crema, Italy, Sept. 16-18, 2002.
- Program Committee Member:
  Int. Conf. on Neural Information Processing, Singapore, Nov. 18-22, 2002
  European Conference on Machine Learning, Helsinki, Finland, Aug. 19-23, 2002
  Int. Workshop on Generative-Model-Based Vision, Copenhagen, Denmark, June 2, 2002.

Dr. Jukka Iivarinen:

- Session Chairman and Program Committee Member, 10th Finnish Artificial Intelligence Conf., Oulu, Finland, Sept. 16-17, 2002.

- Session Chairman, 7th Int. Conf. on Control, Automation, Robotics and Vision. Singapore, Dec. 3-6, 2002
- Program Committee Member, 13th Scandinavian Conf. on Image Analysis, Göteborg, Sweden, June 29 - July 2, 2003.

Dr. Samuel Kaski:

- Editorial Board Member:
  International Journal of Neural Systems (Singapore)
  Intelligent Data Analysis (The Netherlands).
- Member of the Machine Learning for Signal Processing Technical Committee of the IEEE Signal Processing Society, USA (2003-).
- Invited talk "Discriminative clustering: vector quantization in learning metrics." 26th Annual Conf. of the Gesellschaft für Klassifikation (GfKl) July 22-24, 2002, Mannheim, Germany.
- Plenary talk "Learning metrics." 3rd Conf. of the International Society for Ecological Informatics (ISEI), Rome, Italy, Aug. 26-30, 2002.
- Invited talk "Learning metrics." New Trends in Intelligent Information Processing and Web Mining, Zakopane, Poland, June 2-5, 2003.
- Invited talk "Discriminative clustering." 54th Session of the International Statistical Institute (ISI), Berlin, Germany, Aug. 13-20, 2003.
- Plenary talk "Exploration of gene expression." Workshop on Self-Organizing Maps (WSOM'03), Kitakyushu, Japan, Sept. 11-14, 2003.
- Program Chairman and Session Chairman, 10th Finnish Artificial Intelligence Conf., Oulu, Finland, Sept. 16-17, 2002.
- Session Chairman and Program Committee Member:
  9th Int. Conf. on Neural Information Processing (ICONIP'02), Singapore, Nov. 18-22, 2002
  Workshop on Self-Organizing Maps, WSOM'03, Hibikino, Japan, Sept. 11-14, 2003.
- Program Committee Member:
  Int. Symposium on Intelligent Data Engineering and Automated Learning (IDEAL'02), Manchester, U.K., Aug. 12-14, 2002
  Int. Conf. on Data Warehousing and Knowledge Discovery (DaWaK'02), Aix-en-Provence, France, Sept. 4-6, 2002
  IASTED Int. Symposium on Artificial Intelligence and Applications, Innsbruck, Austria, Feb. 18-21, 2002
  6th European Conf. on Principles and Practice of Knowledge Discovery in Databases, Helsinki, Finland, Aug. 19-23, 2002
  Fourth Int. Conf. on Intelligent Data Engineering and Automated Learning (IDEAL'03), March 21-23, 2003, Hong Kong
  5th Int. Conf. on Data Warehousing and Knowledge Discovery, DaWaK 2003, Sept. 3-5, 2003, Prague, Czech Republic
  7th European Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD), Dubrovnik, Croatia, Sept. 22-26, 2003

IIPWM'03, New Trends in Intelligent Information Processing and Web Mining, Zakopane, Poland, June 2-5, 2003
WI 2003, the 2003 IEEE/WIC Int. Conf. on Web Intelligence, Beijing, China, Oct. 13-17, 2003
2003 IEEE Int. Workshop on Neural Networks for Signal Processing, Toulouse, France, Sept. 17-19, 2003.

Dr. Miki Sirola:

- Technical Committee on Modelling and Simulation 2000–2003. Member.
- Technical Committee on Intelligent Systems and Control 2002–2005. Member.
- Session Chairman:
  Int. Conf. on Knowledge-Based Intelligent Information & Engineering Systems (KES'2003), Oxford, U.K., Sept. 3-5, 2003
  IEEE Int. Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS'2003), Lviv, Ukraine, Sept. 8-10, 2003.
- Program Committee Member:
  Int. Conf. on Modelling, Identification and Control, Innsbruck, Austria, Feb. 18-21, 2002
  Int. Conf. on Modelling and Simulation, Marina del Ray, Calif., USA, May 13-15, 2002
  Int. Conf. on Applied Simulation and Modelling, Crete, Greece, June 25-28, 2002
  Int. Conf. on Modelling, Identification and Control, Innsbruck, Austria, Feb. 10-13, 2003
  Int. Conf. on Modelling and Simulation, Palm Springs, USA, Feb. 24-26, 2003
  Int. Conf. on Intelligent Systems and Control, Salzburg, Austria, June 25-27, 2003
  Int. Conf. on Applied Simulation and Modelling, Marbella, Spain, Sept. 3-5, 2003.

Important domestic positions of trust held by researchers of the unit

Academy Professor Erkki Oja:

- Finnish Academy of Science and Letters. Group of mathematics and computer science, vice chairman.
- Opponent at the doctoral dissertation of
  Ville Kyrki, Lappeenranta University of Technology, 2002
  Topi Mäenpää, University of Oulu, 2003.
- Evaluator in filling the academic chairs of professors:
  computer science, University of Helsinki, 2002
  biomedical engineering, University of Oulu, 2003.

Professor Olli Simula:

- Opponent at the doctoral dissertation of
  Lasse Lensu, Lappeenranta University of Technology, 2002
  Olli Saarela, Tampere University of Technology, 2002
  Heikki Jokinen, Tampere University of Technology, 2003.
- Evaluator in filling the academic chair of professor, metrology, Tampere University of Technology, 2002.

Academician Teuvo Kohonen:

- Invited talk "Neural networks." Annual meeting of the union of teachers of mathematical subjects (MAOL ry), Raahe, Finland, Feb. 1, 2002.

Professor Jaakko Hollmén:

- Opponent at the doctoral dissertation of Sampsa Hautaniemi, Tampere University of Technology, 2003.

Professor Timo Honkela:

- Editorial Board Member, journal Puhe ja kieli (Speech and Language).
- Member of IFIP Board, Finnish Information Processing Association.
- Opponent at the doctoral dissertation of Sam Sandqvist, HUT, 2002.
- Evaluator for docentship at Media Lab, University of Art and Design, Helsinki, 2003.

Dr. Jukka Iivarinen:

- Pattern Recognition Society of Finland. Chairman –2002, Vice Chairman 2003– .

Dr. Samuel Kaski:

- Finnish Artificial Intelligence Society. Vice Chairman. 2002.
- Opponent at the doctoral dissertation of
  Vesa Ollikainen, University of Helsinki, 2002
  Jani Mäntyjärvi, University of Oulu, 2003.

Dr. Jorma Laaksonen:

- Finnish Artificial Intelligence Society. Chairman of the dictionary committee. 2003.

Dr. Krista Lagus:

- Member of the Board, Graduate School of Language Technology in Finland
- Member of the Board of FICLA, Finnish Cognitive Linguistics Association

M.Sc. Tapani Raiko:

- Finnish Artificial Intelligence Society. Member of the Governing Board. 2003.

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

- Dr. Patrik Hoyer, New York University, USA, July 2002 (2 wks), Jan. – Dec. 2003.
- Dr. Aapo Hyvärinen, Stanford University, USA, July 2002 (2 wks).
- Dr. Aapo Hyvärinen, Max Planck Institute for Biological Cybernetics, Oct. 2002 (2 wks).
- Dr. Mikko Kurimo, University of Colorado, The Center of Spoken Language Research, Boulder, USA, Dec. 2002–Jan. 2003.
- M.Sc. Tapani Raiko, Albert-Ludwigs-Universität, Freiburg, Germany, Oct. 2001–Sept. 2002.
- Dr. Panu Somervuo, The International Computer Science Institute, Berkeley, USA, Feb. 2002–Jan. 2003.
- Dr. Ricardo Vigário, Fraunhofer Institute for Computer Architecture and Software Technology, Berlin, Germany, Feb.-April, 2002.
- M.Sc. Alexander Ilin, Institut National Polytechnique de Grenoble (INPG), France, May–June 2003.
- Dr. Timo Honkela, Lund University Cognitive Science, Sweden, May–June 2003 (5 wks).
- Dr. Harri Valpola, University of Zürich, Switzerland, Sept. 2003– .

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

- Dr. Mark Plumbley, Queen Mary University of London, U.K., April–July 2002 (3 mths).
- Dr. Maria Funaro, University of Salerno, Italy, Oct. 2002 (2 wks).
- M.Sc. Emilio Di Meglio, University of Napoli, Italy, Feb.-June 2002.
- B.Sc. Leah Russell, Dalarna University, Borlänge, Sweden, June–Nov. 2002.
- M.Sc. Marco Bressan, Universitat Autonoma de Barcelona, Spain, July–Sept. 2002 (6 wks).
- M.Sc. Ana Gonzales, University of La Rioja, Spain, Oct. 2003 (2 wks).
- M.Sc. Patricia Rufino Oliveira, University of Sao Paulo, Brazil, March–Aug. 2003.
- Mr. Jan-Hendrik Schleimer, University of Tübingen, Germany, Aug. 2003– .
- M.Sc. Tanja Kämpfe, University of Bielefeld, Germany, Sept.–Nov. 2003.
- Dr. Ata Kaban, University of Birmingham, U.K., July–Aug. 2003 (6 wks).
- M.Sc. Nasser Mourad, South Valley University, Aswan, Egypt, Oct. 2003– .
- M.Sc. Jose-Miguel Leiva, Universidad Carlos III, Madrid, Spain, Oct.-Nov. 2003.

Courses

Courses given by the Laboratory of Computer and Information Science.

Spring Semester 2002

Code	Course	Lecturer	Course Assistant
T-61.140	Signal Processing Systems	O. Simula	J. Parviainen, J. Pakkanen
T-61.152	Seminar on Computer and Information Science	V. Vuori	
T-61.190	Special Course: *From Gene Expression to Regulation of Cell Function*	J. Hollmén, S. Kaski, H. Wikman	M. Koivisto
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	J. Särelä M. Aksela
T-61.281	Statistical Natural Language Processing	K. Lagus	V. Siivola
T-61.182	Special Course II: *Neural Networks for Speech Processing*	M. Kurimo	V. Siivola
T-61.183	Special Course III: *Biomedical Signal Processing*	R. Vigario	J. Särelä
T-122.102	Special Course VI: *Energy-aware computation*	H. Mannila, P. Orponen	K. Puolamäki

Fall Semester 2002

Code	Course	Lecturer	Course Assistant
T-61.123	Computer Architecture	M. Huttunen	M. Rättö
T-61.124	Special Project in Computer Architecture	M. Huttunen	M. Rättö
T-61.231	Principles of Pattern Recognition	V. Vuori	M. Koskela M. Aksela
T-61.238	Statistical Signal Modelling	P. Pajunen	E. Bingham
T-61.246	Digital Signal Processing and Filtering	O. Simula	J. Parviainen, P. Lehtimäki, T. Similä, V. Viitaniemi, R. Öörni
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	J. Venna
T-61.181	Special Course I: *Time in Self-Organizing Maps*	K. Raivio, O. Simula	P. Lehtimäki
T-61.184	Special Course IV: *Audio Mining*	M. Kurimo	V. Siivola
T-122.101	Special Course V: *Graphical Models*	J. Hollmén, H. Mannila	S. Ruosaari

Spring Semester 2003

Code	Course	Lecturer	Course Assistant
T-61.140	Signal Processing Systems	O. Simula	J. Parviainen, V. Viitaniemi R. Öörni
T-61.152	Seminar on Computer and Information Science	E. Bingham	
T-61.190	Special Course: *From Gene Expression to Regulation of Cell Function*	S. Kaski, J. Hollmén, H. Wikman	M. Oja
T-61.233	Computer Vision	J. Laaksonen	J. Iivarinen
T-61.256	Learning Models and Methods	P. Pajunen	A. Patrikainen
T-61.261	Principles of Neural Computing	K. Raivio	J. Venna M. Aksela
T-61.281	Statistical Natural Language Processing	T. Honkela	V. Siivola
T-61.182	Special Course II: *Robustness in Language and Speech Processing*	M. Kurimo	T. Hirsimäki
T-61.183	Special Course III: *Support Vector Machines and Kernel Methods*	J. Karhunen	K. Raju
T-122.102	Special Course VI: *Analysis of Binary Data*	J. Hollmén, H. Mannila	J. Seppänen

Fall Semester 2003

Code	Course	Lecturer	Course Assistant
T-61.123	Computer Architecture	S. Haltsonen	J. Gröndahl, A. Sorjamaa, J. Takala
T-61.124	Special Project in Computer Architecture	S. Haltsonen	
T-61.140	Signal Processing Systems	O. Simula	J. Parviainen
T-61.231	Principles of Pattern Recognition	T. Honkela K. Raivio	M. Koskela M. Aksela
T-61.238	Statistical Signal Modelling	P. Pajunen	E. Bingham
T-61.246	Digital Signal Processing and Filtering	O. Simula	J. Parviainen, A. Rasinen, V. Viitaniemi
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	J. Venna
T-122.103	Algorithmic methods of data mining	H. Mannila	K. Puolamäki
T-61.184	Special Course IV: *Statistical and Adaptive Approaches to Conceptual Modeling*	T. Honkela K. Lagus J. Särelä	
T-122.101	Special Course V: *Modeling and Mining the Web*	J. Hollmén	A. Patrikainen

Doctoral dissertations

Data exploration process based on the self-organizing map

Juha Vesanto

Dissertation for the degree of Doctor of Science in Technology on 16 May 2002.

External examiners:
Jari Kangas (Nokia Research Center)
Jouko Lampinen (Helsinki University of Technology)
Opponent:
Alfred Ultsch (Philipps-University of Marburg, Germany)

Abstract:

With the advances in computer technology, the amount of data that is obtained from various sources and stored in electronic media is growing at exponential rates. Data mining is a research area which answers to the challange of analysing this data in order to find useful information contained therein. The Self-Organizing Map (SOM) is one of the methods used in data mining. It quantizes the training data into a representative set of prototype vectors and maps them on a low-dimensional grid. The SOM is a prominent tool in the initial exploratory phase in data mining.

The thesis consists of an introduction and ten publications. In the publications, the validity of SOM-based data exploration methods has been investigated and various enhancements to them have been proposed. In the introduction, these methods are presented as parts of the data mining process, and they are compared with other data exploration methods with similar aims.

The work makes two primary contributions. Firstly, it has been shown that the SOM provides a versatile platform on top of which various data exploration methods can be efficiently constructed. New methods and measures for visualization of data, clustering, cluster characterization, and quantization have been proposed. The SOM algorithm and the proposed methods and measures have been implemented as a set of Matlab routines in the SOM Toolbox software library.

Secondly, a framework for SOM-based data exploration of table-format data - both single tables and hierarchically organized tables - has been constructed. The framework divides exploratory data analysis into several sub-tasks, most notably the analysis of samples and the analysis of variables. The analysis methods are applied autonomously and their results are provided in a report describing the most important properties of the data manifold. In such a framework, the attention of the data miner can be directed more towards the actual data exploration task, rather than on the application of the analysis methods. Because of the highly iterative nature of the data exploration, the automation of routine analysis tasks can reduce the time needed by the data exploration process considerably.

Probabilistic models of early vision

Patrik Hoyer

Dissertation for the degree of Doctor of Science in Technology on 15 November 2002.

External examiners:
Mikko Lehtokangas (Tampere University of Technology)
Pentti Laurinen (University of Helsinki)
Opponent:
Eero Simoncelli (University of New York)

Abstract:

How do our brains transform patterns of light striking the retina into useful knowledge about objects and events of the external world? Thanks to intense research into the mechanisms of vision, much is now known about this process. However, we do not yet have anything close to a complete picture, and many questions remain unanswered. In addition to its clinical relevance and purely academic significance, research on vision is important because a thorough understanding of biological vision would probably help solve many major problems in computer vision.

A major framework for investigating the computational basis of vision is what might be called the probabilistic view of vision. This approach emphasizes the general importance of uncertainty and probabilities in perception and, in particular, suggests that perception is tightly linked to the statistical structure of the natural environment. This thesis investigates this link by building statistical models of natural images, and relating these to what is known of the information processing performed by the early stages of the primate visual system.

Recently, it was suggested that the response properties of simple cells in the primary visual cortex could be interpreted as the result of the cells performing an independent component analysis of the natural visual sensory input. This thesis provides some further support for that proposal, and, more importantly, extends the theory to also account for complex cell properties and the columnar organization of the primary visual cortex. Finally, the application of these methods to predicting neural response properties further along the visual pathway is considered.

Although the models considered account for only a relatively small part of known facts concerning early visual information processing, it is nonetheless a rather impressive amount considering the simplicity of the models. This is encouraging, and suggests that many of the intricacies of visual information processing might be understood using fairly simple probabilistic models of natural sensory input.

Unsupervised pattern recognition methods for exploratory analysis of industrial process data

Esa Alhoniemi

Dissertation for the degree of Doctor of Science in Technology on 13 December 2002.

External examiners:
Heikki Hyötyniemi (Helsinki University of Technology)
Jussi Parkkinen (University of Joensuu)
Opponents:
Hannu Koivisto (Tampere University of Technology)
Jussi Parkkinen (University of Joensuu)

Abstract:

The rapid growth of data storage capacities of process automation systems provides new possibilities to analyze behavior of industrial processes. As existence of large volumes of measurement data is a rather new issue in process industry, long tradition of using data analysis techniques in that field does not yet exist. In this thesis, unsupervised pattern recognition methods are shown to represent one potential and computationally efficient approach in analysis of such data.

This thesis consists of an introduction and six publications. The introduction contains a survey on process monitoring and data analysis methods, exposing the research which has been carried out in the fields so far. The introduction also points out the tasks in the process management framework where the methods considered in this thesis – self-organizing maps and cluster analysis – can be benefited.

The main contribution of this thesis consists of two parts. The first one is the use of the existing and development of novel SOM-based methods for process monitoring and data analysis purposes. The second contribution is a concept where cluster analysis is used to extract and identify operational states of a process from measured data. In both cases, the methods have been successfully applied in analysis of real data from processes in the wood processing industry.

Adaptive methods for on-line recognition of isolated handwritten characters

Vuokko Vuori

Dissertation for the degree of Doctor of Science in Technology on 14 December 2002.

External examiners:
Tapio Seppänen (University of Oulu)
Jukka Heikkonen (Helsinki University of Technology)
Opponents:
Tapio Seppänen (University of Oulu)
Louis Vuurpijl (Nijmegen University, Netherlands)

Abstract:

The main goal of the work presented in this thesis has been the development of an on-line handwriting recognition system which is able to recognize handwritten characters of several different writing styles and is able to improve its performance by adapting itself to new writing styles. The recognition method should be applicable to hand-held devices of limited memory and computational resources. The adaptation process should take place during normal use of the device, not in some specific training mode. For the usability aspect of the recognition system, the recognition and adaptation processes should be easily understandable to the users.

The first part of this thesis gives an introduction to the handwriting recognition. The topics considered include: the variations present in personal handwriting styles; automatic grouping of similar handwriting styles; the differences between writer-independent and writer-dependent as well as on-line and off-line handwriting recognition problems; the different approaches to on-line handwriting recognition; the previous adaptive recognition systems and the experiments performed with them; the recognition performance requirements and other usability issues related to on-line handwriting recognition; the current trends in on-line handwriting recognition research; the recognition results obtained with the most recent recognition systems; and the commercial applications.

The second part of the thesis describes an adaptive on-line character recognition system and the experiments performed with it. The recognition system is based on prototype matching. The comparisons between the character samples and prototypes are based on the Dynamical Time Warping (DTW) algorithm and the input characters are classified according to the k Nearest Neighbors (k-NN) rule. The initial prototype set is formed by clustering character samples collected from a large number of subjects. Thus, the recognition system can handle various writing styles. This thesis work introduces four DTW-based clustering algorithms which can be used for the prototype selection. The recognition system adapts to new writing styles by modifying its prototype set. This work introduces several adaptation strategies which add new writer-dependent prototypes into the initial writer-independent prototype set, reshape the existing prototypes with a Learning Vector Quantization (LVQ)-based algorithm, and inactivate poorly performing prototypes. The adaptations are carried out on-line in a supervised or self-supervised fashion. In the former case, the user explicitly labels the input characters which are used

as training samples in the adaptation process. In the latter case, the system deduces the labels from the recognition results and the user's actions. The latter approach is prone to erroneously labeled learning samples.

The different adaptation strategies were experimented with and compared with each other by performing off-line simulations and genuine on-line user experiments. In the simulations, special attention has been paid to the various erroneous learning situations likely to be encountered in real world handwriting recognition tasks. The recognition system is able to improve its recognition accuracy significantly on the basis of only a few additional character samples per class. Recognition accuracies acceptable in real world applications can be attained for most of the test subjects.

This work also introduces a Self-Organizing Map (SOM)-based method for analyzing personal writing styles. Personal writing styles are represented by high-dimensional vectors, the components of which indicate the subjects' tendencies to use certain prototypical writing styles for isolated characters. These writing style vectors are then visualized by a SOM which enables the detection and analysis of clusters of similar writing styles.

Using visualization, variable selection and feature extraction to learn from industrial data

Sampsa Laine

Dissertation for the degree of Doctor of Science in Technology on 19 September 2003.

External examiners:
Olli Saarela (Keskuslaboratorio Oy)
Petri Vasara (Jaakko Pöyry Group Oyj)
Opponent:
John Klus (University of Wisconsin-Madison)

Abstract:

Although the engineers of industry have access to process data, they seldom use advanced statistical tools. Why this reluctance? I believe engineers do not have adequate statistical skills, and that inexpert use of statistics leads to useless results. For example, failure to correctly identify and remove outliers disturbs those tools that assume Gaussian distribution of data. Also, failure to correctly parameterize the used algorithm leads to poor results, as an example, a process engineer may find it difficult to find the best structure of an artificial neural network. Failures of statistical tools lead the engineer to disregard statistics, and resort to visual study of manually selected data.

This thesis looks for algorithms that serve the common process engineer. I prefer three properties in an algorithm: supervised operation, robustness and understandability. Supervised operation allows and requires the user to explicate the goal of the analysis, which allows the algorithm to discover results that are relevant to the user. Robust algorithms allow engineers to analyse raw process data collected from the automation system of the plant. Understandability is the most important criterion: the user must understand how to parameterize the model, what is the principle of the algorithm, and know how to interpret the results.

These criteria are used to assess algorithms for visualization, variable selection and feature extraction. The objective of this thesis was to create a tool set the reliably and understandably provides the user with information that is related to a problem that he/she has defined interesting.

The tools and the criteria are illustrated by analysing an industrial case: the concentrator of the Hitura mine. This case illustrates how to define the problem using off-line laboratory data; and how to study the on-line data to find solutions. Statistical tools demonstratedly improve the efficiency of process study: my early results required approximately six man months of work; the algorithms proposed by this thesis produced comparable results in few weeks.

Interactive image retrieval using self-organizing maps

Markus Koskela

Dissertation for the degree of Doctor of Science in Technology on 14 November 2003.

External examiners:
Irwin King (Chinese University of Hong Kong)
Timo Ojala (University of Oulu)
Opponent:
Moncef Gabbouj (Tampere University of Technology)

Abstract:

Digital image libraries are becoming more common and widely used as visual information is produced at a rapidly growing rate. Creating and storing digital images is nowadays easy and getting more affordable all the time as the needed technologies are maturing and becoming eligible for general use. As a result, the amount of data in visual form is increasing and there is a strong need for effective ways to manage and process it. In many settings, the existing and widely adopted methods for text-based indexing and information retrieval are inadequate for these new purposes.

Content-based image retrieval addresses the problem of finding images relevant to the users' information needs from image databases, based principally on low-level visual features for which automatic extraction methods are available. Due to the inherently weak connection between the high-level semantic concepts that humans naturally associate with images and the low-level visual features that the computer is relying upon, the task of developing this kind of systems is very challenging. A popular method to improve retrieval performance is to shift from single-round queries to navigational queries where a single retrieval instance consists of multiple rounds of user-system interaction and query reformulation. This kind of operation is commonly referred to as relevance feedback and can be considered as supervised learning to adjust the subsequent retrieval process by using information gathered from the user's feedback.

In this thesis, an image retrieval system named PicSOM is presented, including detailed descriptions of using multiple parallel Self-Organizing Maps (SOMs) for image indexing and a novel relevance feedback technique. The proposed relevance feedback technique is based on spreading the user responses to local SOM neighborhoods by a convolution with a kernel function. A broad set of evaluations with different image features, retrieval tasks, and parameter settings demonstrating the validity of the retrieval method is described. In particular, the results establish that relevance feedback with the proposed method is able to adapt to different retrieval tasks and scenarios.

Furthermore, a method for using the relevance assessments of previous retrieval sessions or potentially available keyword annotations as sources of semantic information is presented. With performed experiments, it is confirmed that the efficiency of semantic image retrieval can be substantially increased by using these features in parallel with the standard low-level visual features.

Learning metrics and Discriminative Clustering

Janne Sinkkonen

Dissertation for the degree of Doctor of Philosophy on 21 November 2003.

External examiners:
Petri Myllymäki (University of Helsinki)
Kari Torkkola (Motorola)
Opponent:
Naftali Tisby (Hebrew University of Jerusalem)

Abstract:

In this work methods have been developed to extract relevant information from large, multivariate data sets in a flexible, nonlinear way. The techniques are applicable especially at the initial, explorative phase of data analysis, in cases where an explicit indicator of relevance is available as part of the data set.

The unsupervised learning methods, popular in data exploration, often rely on a distance measure defined for data items. Selection of the distance measure, part of which is feature selection, is therefore fundamentally important.

The learning metrics principle is introduced to complement manual feature selection by enabling automatic modification of a distance measure on the basis of available relevance information. Two applications of the principle are developed. The first emphasizes relevant aspects of the data by directly modifying distances between data items, and is usable, for example, in information visualization with the self-organizing maps. The other method, discriminative clustering, finds clusters that are internally homogeneous with respect to the interesting variation of the data. The techniques have been applied to text document analysis, gene expression clustering, and charting the bankruptcy sensitivity of companies.

In the first, more straightforward approach, a new local metric of the data space measures changes in the conditional distribution of the relevance-indicating data by the Fisher information matrix, a local approximation of the Kullback-Leibler distance. Discriminative clustering, on the other hand, directly minimizes a Kullback-Leibler based distortion measure within the clusters, or equivalently maximizes the mutual information between the clusters and the relevance indicator. A finite-data algorithm for discriminative clustering is also presented. It maximizes a partially marginalized posterior probability of the model and is asymptotically equivalent to maximizing mutual information.

Computational models relating properties of visual neurons to natural stimulus statistics

Jarmo Hurri

Dissertation for the degree of Doctor of Science in Technology on 5 December 2003.

External examiners:
Pentti Laurinen (University of Helsinki)
Jukka Heikkonen (Helsinki University of Technology)
Opponent:
Laurenz Wiskott (Humboldt-Universität zu Berlin)

Abstract:

The topic of this thesis is mathematical modeling of computations taking place in the visual system, the largest sensory system in the primate brain. While a great deal is known about how certain visual neurons respond to stimuli, a very profound question is *why* they respond as they do. Here this question is approached by formulating models of computation which might underlie the observed response properties. The main motivation is to improve our understanding of how the brain functions. A better understanding of the computational underpinnings of the visual system may also yield advances in medical technology or computer vision, such as development of visual prostheses, or design of computer vision algorithms.

In this thesis several models of computation are examined. An underlying assumption in this work is that the statistical properties of visual stimuli are related to the structure of the visual system. The relationship has formed through the mechanisms of evolution and development. A model of computation specifies this relationship between the visual system and stimulus statistics. Such a model also contains free parameters which correspond to properties of visual neurons. The experimental evaluation of a model consists of estimation of these parameters from a large amount of natural visual data, and comparison of the resulting parameter values against neurophysiological knowledge of the properties of the neurons, or results obtained with other models.

The main contribution of this thesis is the introduction of new models of computation in the primary visual cortex. The results obtained with these models suggest that one defining feature of the computations performed by a class of neurons called simple cells, is that the output of a neuron consists of periods of intense neuronal activity. It also seems that the activity levels of nearby simple cells are positively correlated over short time intervals. In addition, the probability of the occurrence of such regions of intense activity in the joint space of time and cortical area seems to be small. Another contribution of the thesis is the examination of the relationship between two previous computational models, namely independent component analysis and local spatial frequency analysis. This examination suggests that results obtained with independent component analysis share some important properties with wavelets, in the way their localization in space and frequency depends on their average spatial frequency.

Advances in Independent Component Analysis with applications to data mining

Ella Bingham

Dissertation for the degree of Doctor of Science in Technology on 12 December 2003.

External examiners:
Thomas Hoffman (Brown University)
Helena Ahonen-Myka (University of Helsinki)
Opponent:
Mark Plumbley (Queen Mary University of London)

Abstract:

This thesis considers the problem of finding latent structure in high dimensional data. It is assumed that the observed data are generated by unknown latent variables and their interactions. The task is to find these latent variables and the way they interact, given the observed data only. It is assumed that the latent variables do not depend on each other but act independently.

A popular method for solving the above problem is independent component analysis (ICA). It is a statistical method for expressing a set of multidimensional observations as a combination of unknown latent variables that are statistically independent of each other. Starting from ICA, several methods of estimating the latent structure in different problem settings are derived and presented in this thesis. An ICA algorithm for analyzing complex valued signals is given; a way of using ICA in the context of regression is discussed; and an ICA-type algorithm is used for analyzing the topics in dynamically changing text data. In addition to ICA-type methods, two algorithms are given for estimating the latent structure in binary valued data. Experimental results are given on all of the presented methods.

Another, partially overlapping problem considered in this thesis is dimensionality reduction. Empirical validation is given on a computationally simple method called random projection: it does not introduce severe distortions in the data. It is also proposed that random projection could be used as a preprocessing method prior to ICA, and experimental results are shown to support this claim.

This thesis also contains several literature surveys on various aspects of finding the latent structure in high dimensional data.

Theses

Licentiate of Science in Technology

2002

Ilvonen, Mikko
Constrained optimization of the VALMA dose assessment model by a genetic algorithm

2003

Juutila, Rauno
Jalkapallon syöttöjakaumien visualisointi, ryhmittely ja SOM-kartoitus (Visualization, grouping and SOM-mapping of football's passing distribution)

Master of Science in Technology

2002

Hirsimäki, Teemu
A decoder for large-vocabulary continuous speech recognition

Häkkinen, Marko
Developing new features through reengineering of legacy software

Karanko, Markus
Extrapolating precipitation area motion from a sequence of radar images

Katajamaa, Mikko
Simulation model for exploring variations in the gene expression data and its analysis

Kurhila, Mikko
The study of soft-decision algorithms for the TETRA base station system

Kuusisto, Jukka
Puheenaiheiden tunnistus puheenvuoroista oppivilla menetelmillä (Recognition of dialogue topics with learning methods)

Laakso, Harri
Communication system in a digital signal processing software platform

Lehtimäki, Pasi
Self-organizing operator maps in complex system analysis

Oja, Merja
Geeniekspressiotiedon louhinta itseorganisoivien karttojen ja oppivien metriikoiden avulla (Gene expression data mining using self-organizing maps and learning metrics)

Pakkanen, Jussi
Sisältöpohjainen haku paperivirhetietokannassa PicSOM-järjestelmän avulla (Content-based retrieval in a paper defect database using the PicSOM system)

Patrikainen, Anne
Projected clustering of high-dimensional binary data

Repo, Juha
Quality monitoring and visualisation platform for road construction

Saari, Aleksi
Topographic mappings for analyzing clinical patient data

Sarmavuori, Juha
Kaiunpoisto ja tandemkoodauksen esto (Echo cancellation and tandem free operation)

Savolainen, Liina
SAR image projecting and visualization & analysis tool for sea ice research

Ursin, Markku
Triphone clustering in Finnish continuous speech recognition

2003

Aho, Ilkka
Epävarmuustekijöiden huomioon ottaminen päästökaupan päätöksenteon simulointimallissa (Uncertainty factors in the simulation model of emissions trading and decision-making)

Juntunen, Heli
Finding components in discrete biosequences

Klami, Arto
Regularized discriminative clustering

Kostiainen, Jukka
A data warehousing solution to paper mill production and quality reporting

Lahti, Leo
Vertaileva toiminnallinen genomianalyysi assosiatiivisella ryhmittelymenetelmällä (Comparative functional genome analysis using associative clustering)

Laurila, Jouni
Sähkönmyyjän riskianalyysin tarpeet ja tekninen suunnittelu (Needs and technical design of electricity retailer's risk analysis)

Lindström, Jukka
An architectural framework for data mining

Malkki, Jussi
Prediction of absorption behaviour of a drug with in vitro data using compartment model in Bayesian data analysis

Minkkinen, Sami
Diabeteksen MODY3-fenotyypin todennäköisyyden estimointi ja visualisointi (Probability estimation and visualization of the MODY3 phenotype of diabetes)

Paatero, Vesa
Sanojen painotusmenetelmien vertailu WEBSOM-kartoilla (A comparison of term weighting methods using WEBSOM maps)

Rummukainen, Mika
Implementing multimedia retrieval markup language for image retrieval systems' comparison

Setälä, Henri
Automatic fiducial detection from radiographs

Siekkinen, Matti
Aspect oriented make

Sulkava, Mikael
Identifying spatial and temporal profiles from forest nutrition data

Syrjälä, Juha
Context classification using audio data for a wearable computer

Viiperi, Sampo
Feature selection for the purpose of segmentation of polysomnograms in subjects with developmental brain disorders

Viitaniemi, Ville
Image segmentation in content-based image retrieval

I Neural Networks Research Centre Research Projects

Chapter 1

Introduction

Erkki Oja, Director of the Neural Networks Research Centre

The core area of research in the Neural Networks Research Centre is neurocomputing, in the sense it is understood today. 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 early 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 NNRC today.

Pattern analysis and statistical machine learning are the central tools for structuring raw information in the new knowledge economy. Information will need 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. The new insight is

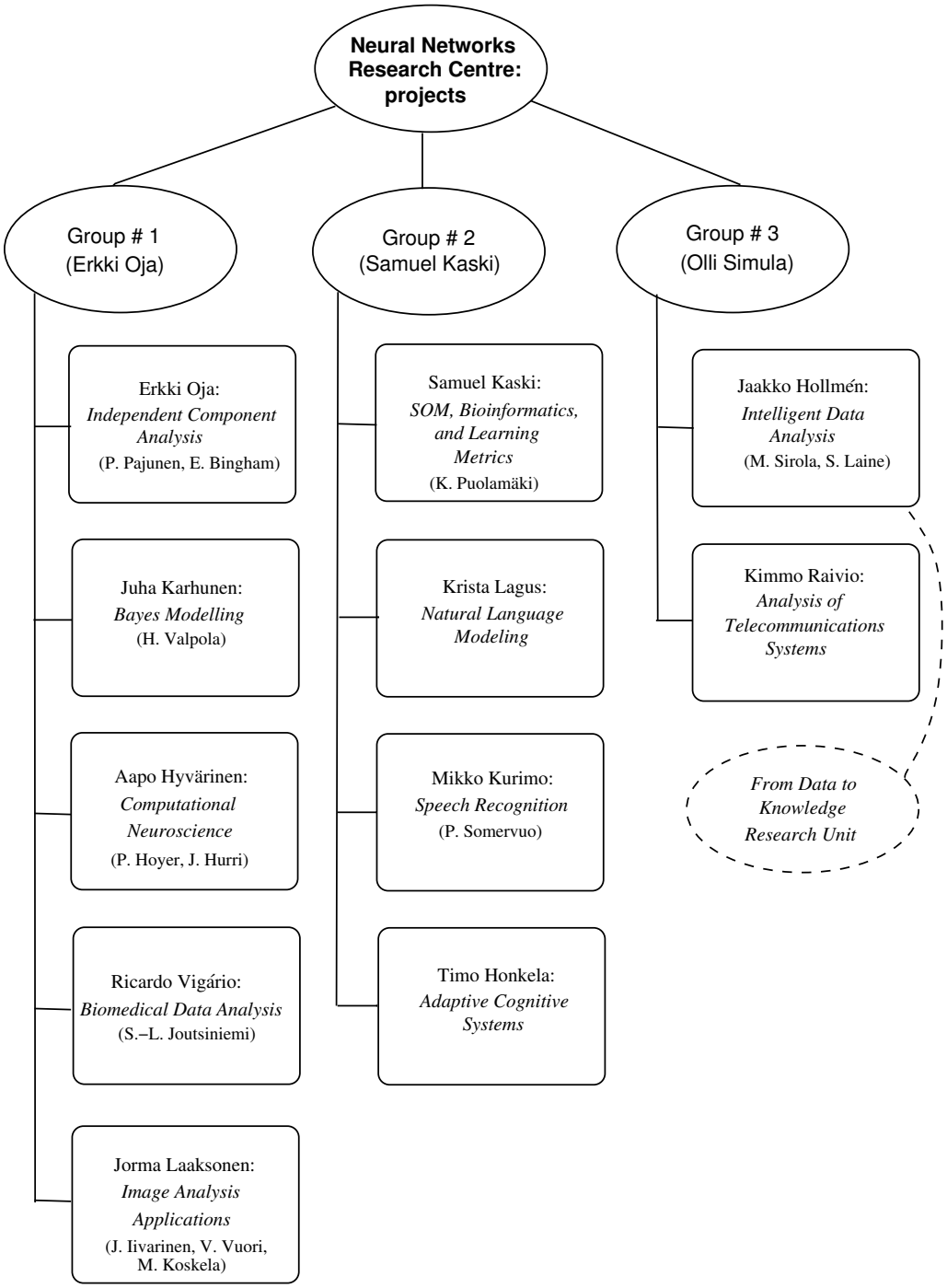


Figure 1.1: The Neural Networks Research Centre consists of three major groups, each having a number of smaller project groups. The leader of each group and the research topic are marked within each box, as well as the names of the post-doctoral researchers within each project. The dotted line indicates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

that although the models are not available in closed form, the intrinsic latent features or factors of the observations, and their mutual interrelations, can be *learned* from the data

using a great number of simultaneously co-operating simple processing units or operations. This is in analogy with the operating principles of the biological neural systems.

In the Neural Networks 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 is a concise description of our internal project organization at the moment. The Research Unit consists 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 (Jan. 2004) is 23, 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, `http://www.cis.hut.fi/research/`.

Chapter 2

Independent component analysis and blind source separation

Erkki Oja, Juha Karhunen, Ella Bingham, Maria Funaro, Johan Himberg, Antti Honkela, Aapo Hyvärinen, Alexander Ilin, Karthikesh Raju, Tapani Ristaniemi, Jaakko Särelä, Harri Valpola, Ricardo Vigário

2.1 Introduction

Erkki Oja

What is Independent Component Analysis? 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.

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 2001. 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 3500 copies have been sold by August, 2003. The book has been extensively cited in the ICA literature and seems to have evolved into the standard text on the subject worldwide. Another tangible contribution has been the FastICA software package (`http://www.cis.hut.fi/projects/ica/fastica/`) which during the reporting period was downloaded by 8900 registered users. 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 2002 - 2003, ICA research stayed as a core project in the laboratory. It was extended to several new directions. It is no more possible to report all this work under a single ICA Chapter. The most advanced developments are now presented in their separate Chapters in this report. They are: "Variational Bayesian learning of generative models" (a project led by prof. Juha Karhunen and Dr. Harri Valpola), "Analysis of independent components in biomedical signals" (a project led by Dr. Ricardo Vigario) and "Computational neuroscience" (a project led by Doc. Aapo Hyvärinen).

This Chapter starts by introducing some theoretical advances undertaken during the reporting period. Also comparisons on post-nonlinear mixtures are reported. Then, several smaller application-oriented ICA projects are covered. The applications range from text mining and astronomical data analysis to telecommunications. Note that more extensive applications are covered especially in the Chapter on biomedical signal analysis. Finally, in the present Chapter, the EU project BLISS is reviewed, which ended during 2003 with highly favorable reviews.

2.2 Theoretical advances

Erkki Oja, Ella Bingham, Aapo Hyvärinen, Jaakko Särelä, Harri Valpola

ICA in a regression problem

In this project it was shown how independent component analysis (ICA) can be used in the context of regression. In a regression problem, one has a set of predictor variables and a set of predicted variables. The task is to generate a mapping between these sets so that given the values of the predictor variables, the values of the predicted variables can be estimated.

The regression problem can be cast into the ICA framework as follows. Using the training data of predictor and predicted variables, the independent components in the data are estimated. Using them we can estimate future values of predicted variables, given the observations of the predictor variables only. The problem is discussed in detail in [1], where it is also shown that regression by ICA is closely related to regression by a multilayer perceptron (MLP) network, which is a widely used neural network. However, regression by ICA is more straightforward and better defined in terms of choosing the number of units and the nonlinear transformations in the hidden layer of the MLP network, and finding the weights of the layers of the network.

Non-negative ICA

The basic linear ICA model can be considered to be solved, with a multitude of practical algorithms and software. However, if one makes some further assumptions which restrict or extend the model, then there is still ground for new theoretical analysis and solution methods. One such assumption is *positivity or non-negativity* of the sources and perhaps the mixing coefficients. Non-negativity is a very natural condition for many practical real-world applications, for example in the analysis of images, text, or spectral data. The constraint of non-negative sources, perhaps with an additional constraint of non-negativity on the mixing matrix, is often known as *positive matrix factorization* or *non-negative matrix factorization*. We refer to the combination of non-negativity and independence assumptions on the sources as *non-negative independent component analysis*.

It was suggested by the co-author of [3] that a suitable cost function for actually finding the rotation could be constructed as follows: denote the estimates for the positive sources by $y_i,\ i = 1, ..., n$, and suppose we have an output truncated at zero, $\mathbf{y}^+ = (y_1^+, ..., y_n^+)$ with $y_i^+ = \max(0, y_i)$. Let us construct a reestimate of whitened but not zero-mean observation data $\mathbf{z} = \mathbf{W}^T\mathbf{y}$, given by $\hat{\mathbf{z}} = \mathbf{W}^T\mathbf{y}^+$, where $\mathbf{W}$ is the (orthogonal) parameter matrix of the ICA problem. Then a suitable cost function would be

$$J(\mathbf{W}) = E\{\|\mathbf{z} - \hat{\mathbf{z}}\|^2\} = E\{\|\mathbf{z} - \mathbf{W}^T\mathbf{y}^+\|^2\} \tag{2.1}$$

because obviously its value will be zero if $\mathbf{W}$ is such that all the y_i are positive, or $\mathbf{y} = \mathbf{y}^+$.

We have considered the minimization of this cost function by on-line learning algorithms. In [3], the cost function (2.1) was taken as a special case of "nonlinear PCA" for which an algorithm was earlier suggested by one of the authors [2]. However, a rigorous convergence proof for the nonlinear PCA method could not be constructed except in some special cases. The general convergence seems a very challenging problem.

In another paper [4] we showed that the cost function (2.1) has very desirable properties in the Stiefel manifold of rotation (orthogonal) matrices: the function has no local minima and it is a Lyapunov function for its gradient matrix flow. A gradient algorithm, suggested

in the paper, is therefore monotonically converging and is guaranteed to find the absolute minimum of the cost function. The minimum is zero, giving positive components y_i, which must be a positive permutation of the original unknown sources s_j, as proven in the paper. Some preliminary results along these lines were given by the authors in [5].

Semi-blind source separation by denoising

Many algorithms have been invented to perform blind source separation. There, it is assumed that the sources are completely unknown. However, in many cases of time series separation there exists some prior knowledge on the behaviour of the sources. This is the case especially, when the data is collected in a controlled experiment.

In [8] we have introduced a fast semi-blind source separation algorithm (SBSS) that allows an easy incorporation of such prior knowledge. In one iteration of SBSS, the essential step denoises the current source estimate from part of the noise as well as the interference of the other sources. Then, the projection is sought that gives a source estimate closest to this denoised source estimate in least mean squares sense.

SBSS introduces a full continuum of algorithms, where the denoising can vary from a very detailed matched filter to looser, more general denoising principles, such as the non-Gaussianity in ICA.

Overlearning in ICA

The research on overlearning in ICA has been continued. We have elaborated the discussion as well as suggested several solutions to solve both the problems of spikes and bumps. We have published a comprehensive article on the findings [6]. Further study on a Bayesian approach to overcome the problem has been conducted as well [7].

References

[1] Aapo Hyvärinen and Ella Bingham. Connection between multilayer perceptrons and regression using independent component analysis. *Neurocomputing*, 50(C):211–222, January 2003.

[2] Oja, E.: The nonlinear PCA learning rule in Independent Component Analysis. *Neurocomputing 17*, pp. 25 - 45 (1997)

[3] Plumbley, M. and Oja, E.: A "Non-negative PCA" algorithm for independent component analysis. *IEEE Trans. on Neural Networks 15*, no. 1, pp. (2004)

[4] Oja, E. and Plumbley, M.: Blind separation of positive sources by globally convergent gradient search. *Neural Computation*, to appear.

[5] Oja, E. and Plumbley, M.: Blind separation of positive sources using non-negative PCA. *Proc. 4th Int. Symp. on Independent Component Analysis and Blind Source Separation*, April 1 - 4, 2003, Nara, Japan, pp. 11 - 16 (2003).

[6] J. Särelä and R. Vigário, "Overlearning in marginal distribution-based ICA: analysis and solutions," *Journal of machine learning research*, vol. 4 (Dec), pp. 1447–1469, 2003.

[7] J. Särelä and R. Vigário. A Bayesian approach to overlearning in ICA. Tech. Rep A 70, Lab of Computer and Information Science, Helsinki University of Technology, Finland, 2003.

[8] H. Valpola and J. Särelä. A fast semi-blind source separation algorithm. Tech. Rep A 66, Lab of Computer and Information Science, Helsinki University of Technology, Finland, 2002.

2.3 Comparison studies on blind separation of post-nonlinear mixtures

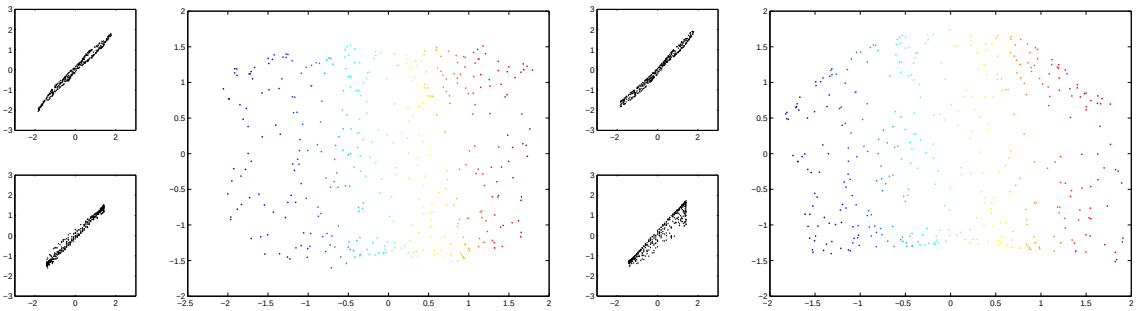
Alexander Ilin, Juha Karhunen

Different approaches proposed for nonlinear independent component analysis (ICA) and blind source separation (BSS) have been recently reviewed in [1]. However, their limitations and domains of preferable application have been studied only a little, and there do not exist hardly any comparisons of the proposed methods. 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 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. The sources were a sine wave and uniformly distributed white noise.

Figure 2.1 shows some of the experimental results. The INPG method based on the independence criterion obtained a good signal-to-noise ratio for the recovered sources. However, it failed to cope with non-invertible post-nonlinearities in the mixtures. The more general Bayesian NFA+FastICA approach was able to recover the sources as well. Moreover, it was able to process mixtures with non-invertible distortions.



(a) (b) (a) (b)

PNL method of INPG: SNR 17.73 dB Bayesian NLFA+FastICA: SNR 13.57 dB

Figure 2.1: The two sources recovered from their PNL mixture by the two alternative approaches: (a) – the scatter plots showing the values of one original source signal plotted against the found sources, (c) – the distribution of the found sources. The INPG method used only mixtures with invertible post-nonlinear distortions. The Bayesian approach was able to process the mixture with one non-invertible post-nonlinearity.

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:

- The INPG methods perform better in PNL mixtures with the same number of sources and observed mixtures when all the post-nonlinearities are invertible.
- The performance of both methods can be improved by exploiting more mixtures than the number of sources.
- The advantage of the Bayesian methods in post-nonlinear BSS problems is that they can separate overdetermined post-nonlinear mixtures with non-invertible post-nonlinearities while the existing INPG methods cannot do this.

The results of this comparison study will be presented in a forthcoming international joint paper.

References

[1] C. Jutten and J. Karhunen. Advances in nonlinear blind source separation. In *Proc. of the 4th Int. Symp. on Independent Component Analysis and Blind Signal Separation (ICA2003)*, pages 915–920, Nara, Japan, 2003. Invited paper in the special session on nonlinear ICA and BSS.

[2] A. Hyvärinen and E. Oja. A Fast Fixed-Point Algorithm for Independent Component Analysis. *Neural Computation*, 9(7):1483–1492, 1997.

[3] Harri Lappalainen and Antti 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.

2.4 Text mining

Ella Bingham

Independent component analysis (ICA) was originally developed for signal processing applications. Recently it has been found out that ICA is a powerful tool for analyzing text document data as well, if the text documents are presented in a suitable numerical form. This opens up new possibilities for automatic analysis of large textual data bases: finding the topics of documents and grouping them accordingly.

First approaches of using ICA in the context of text data considered the data static. In our recent study, we concentrated on text data whose topic changes over time. Examples of dynamically evolving text are chat line discussions or newsgroup documents. The dynamical text stream can be seen as a time series, and methods of time series processing may be used to extract the underlying characteristics — here the topics — of the data.

The project is described in detail in [1]. As an example of dynamically changing textual data, we used chat line discussions where several discussions are going on simultaneously, and the topics of the discussions change dynamically as participants enter and leave the chat room. We were able to find meaningful topics which can be visualized both by different sets of keywords for each topic, and by their behaviour through time (some topics are more or less persistent during the whole period of time; some topics die but will come up again later; and some topics are only active at a certain period of time).

To conclude, our method finds meaningful topics inherent in the data, and the experimental results suggest the applicability of the method to query-based retrieval from a temporally changing text stream.

References

[1] Ella Bingham, Ata Kabán, and Mark Girolami. Topic identification in dynamical text by complexity pursuit. *Neural Processing Letters*, 17(1):69–83, 2003.

2.5 ICA for astronomical data

Erkki Oja, Harri Valpola, Maria Funaro

One of the main research directions in modern astrophysics is to understand the dark matter in the universe, especially the baryonic component, supposed to be formed by compact objects with substellar mass called Massive Astrophysical Compact Halo Objects (MACHO). Possible candidates include small black holes, dwarf stars, or exoplanets. When such an object passes near the line of sight of a star, the luminosity of the star will increase – an effect called gravitational lensing, predicted by the general theory of relativity. In studying other galaxies than our own, individual stars cannot be resolved, but a whole group of unresolved stars is registered in a single pixel element of a telescope CCD camera. In a new technique called pixel lensing, the pixel luminosity variations over time are monitored, and using these time series the lensing events can be detected even in the case of unresolved stars.

A severe problem in the analysis of the images and luminosity variations is the presence of artefacts. One possible cause for artefacts are the individual stars between the far-out galaxy and the camera, which emerge sharply from the luminosity background. Other artefacts are cosmic rays, atmospheric events, and noise in the CCD camera. Separating these artefacts from the interesting astrophysical events is one of the necessary steps in the analysis of pixel lensing data. Artefact removal techniques have to be fast and highly acurate to avoid the interesting phenomena from being eliminated from the data along with the artefacts.

The new idea proposed by us [1] is to use Independent Component Analysis for artefact detection and removal. We can assume that the astrophysical images consist of three additive components: first, images revealing the interesting physical effects like pixel lensing; second, images of artefacts such as atmospheric events, cosmic rays, or the resolved stars; and third, additive noise mainly due to the camera system. All of these are guaranteed to be independent, so the ICA model holds very well. In the image processing that we propose here, we first apply PCA to the images. This has the effect of strongly reducing the additive noise, leaving the images with physical effects and artefacts. Next, we apply ICA on these images, with the result that the artefacts are separated and can be removed. What is left are clean artefact-free images that can be analyzed further for the possible physical phenomena.

Using image data on the M31 Galaxy, it was shown in [1] that several clear artefacts can be detected and recognized based on their temporal pixel luminosity profiles and independent component images. Once these are removed, it is possible to concentrate on the real physical events like gravitational lensing.

References

[1] Funaro, M., Oja, E. and Valpola, H.: Independent component analysis for artefact separation in astrophysical images. *Neural Networks 16*, no. 3-4, pp. 469-478 (2003).

2.6 ICA in CDMA communications

Karthikesh Raju, Tapani Ristaniemi, Juha Karhunen, 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. The project started with application of ICA and BSS methods to various problems in multiuser detection [1, 2], trying to take into account the available prior information whenever possible. We showed that ICA based methods can yield considerably better performances than more conventional methods based on second-order statistics. The work carried out during this stage is reviewed together with the necessary background in Chapter 23 of the book [3].

In the second stage of the project in 2001-2003, we have applied independent component analysis to blind suppression of various interfering signals appearing in direct sequence CDMA communication systems. First we studied bit-pulsed jamming, which constitutes an important problem in practical CDMA communication systems. We have taken into account both data modulation and temporally uncorrelated jamming, improving and extending earlier preliminary work on the same problem. Computer simulations show that the proposed method performs better than the well-known RAKE method, which is the standard choice for suppressing jammer signals. The results have been reported in more detail in the conference papers [4, 5].

In papers [6, 7], ICA-RAKE Pre-Switch and ICA-RAKE Post-Switch structures were introduced. They switch between the ICA portion and the RAKE portion depending on the signal-to-jammer ratio. If the jammer signal is weak or absent, preprocessing by ICA is not advisable, because it might even cause additional interference.

Fig. 2.2 shows the distribution of correct bits using the post-switched ICA-RAKE and plain RAKE methods for a coherent 5 path channel. In the case of a single path (left subfigures), post-switched ICA (upper left subfigure) is able to separate the jammer completely since about 95% of the blocks are correct, while the results provided by the conventional RAKE receiver (shown in the lower left subfigure) are poor. The situation is qualitatively similar in the case of 5 paths, as shown by the right subfigures of Fig. 2.2. These results are summarized in the paper [10].

We have applied ICA also to cancellation of interferences due to adjacent cells in a DS-CDMA system. The gain in performance is about $5-8$dB when the interfering source

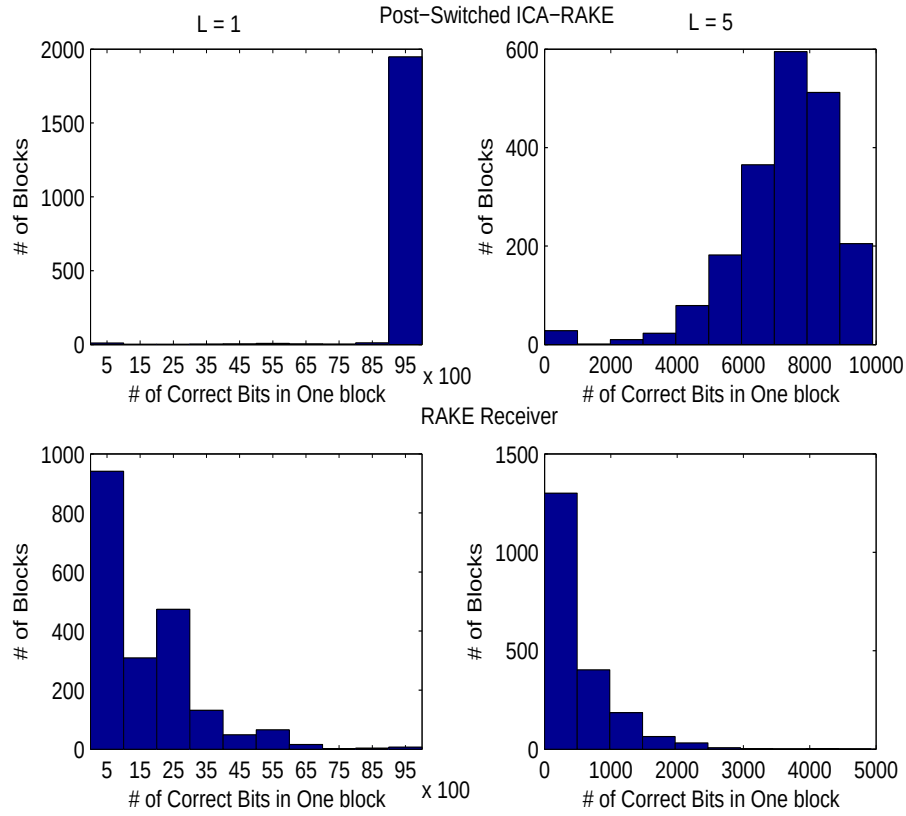


Figure 2.2: Distribution of correct bits per block for post-switched ICA-RAKE (upper subfigures) and plain RAKE (lower subfigures) methods for $L = 1$ (left subfigures) path and $L = 5$ (right subfigures) paths. ICA-RAKE has more than 95% correct bits in most blocks for $L = 1$ while most blocks have at least 70% correct symbols under $L = 5$ coherent paths.

has been suppressed with ICA [8, 9].

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] T. Ristaniemi, K. Raju, and J. Karhunen, Jammer mitigation in DS-CDMA array systems using independent component analysis. In *Proc. of the 2002 IEEE Int. Conf. on Communications (ICC2002)*, New York City, NY, USA, April 28–May 2, 2002.

[5] K. Raju, T. Ristaniemi, J. Karhunen, and E. Oja, Suppression of bit-pulsed jammer signals in DS-CDMA array systems using independent component analysis. In *Proc. of the 2002 IEEE Int. Symp. on Circuits and Systems (ISCAS2002)*, Phoenix, Arizona, USA, May 26-29, 2002, pp. I-189/I-192.

[6] K. Raju and T. Ristaniemi, ICA-RAKE switch for jammer cancellation in DS-CDMA array systems. In *Proc. of the 2002 IEEE Int. Symp. on Spread Spectrum Techniques and Applications (ISSSTA2004)*, Prague, Czech Republic, September 2-5, 2002, pp 638-642.

[7] T. Ristaniemi, K. Raju, J. Karhunen, and E. Oja, Jammer cancellation in DS-CDMA arrays: pre and post switching of ICA and RAKE. In *Proc. of the 2002 IEEE Workshop on Neural Networks for Signal Processing (NNSP2002)*, Martigny, Switzerland, September 4-6, 2002, pp. 495–504.

[8] T. Ristaniemi, K. Raju, and J. Karhunen, Inter-cell interference cancellation in CDMA array systems by independent component analysis. In *Proc. of the 4th Int. Symp. on Independent Component Analysis and Blind Signal Separation (ICA2003)*, Nara, Japan, April 1-4, 2003, pp. 739–744.

[9] K. Raju and T. Ristaniemi, Exploiting independences to cancel interferences due to adjacent cells in a DS-CDMA system. In *Proc. of the Personal, Indoor, and Mobile Radio Communications (PIMRC 2003)*, Beijing, China, September 7-10, 2003.

[10] K. Raju, T. Ristaniemi, and J. Karhunen, Semi-blind interference suppression on coherent multipath environments. Submitted to a conference.

2.7 Explorative investigation of the reliability of independent component estimates

Johan Himberg, Aapo Hyvärinen

FastICA [2] is a fast, fixed point algorithm for estimating independent components[1]. However, FastICA, as most ICA algorithms, finds a local minimum of its objective function. Even with algorithms which are deterministic and always find the global optimum of their objective function, the valid interpretation of the results need some analysis of the statistical reliability or significance of the components. There are two different reasons for this. Firstly, as real data never exactly follows the ICA model, the contrast function used in the estimation may have many local minima which are all equally good. The independent components are simply not well-defined in this case. Secondly, even in the extreme where the data is exactly generated according to the ICA model, the finite sample size induces statistical errors in the estimation—this is the case where classical analysis of statistical significance and confidence intervals would be needed.

As such, the bootstrapping method for analyzing the statistical reliability of independent components in [3] seems applicable only in the case of deterministic algorithms. Some additional development is required for stochastic algorithms, and consequently, we have developed an interactive visualization method and software package[2] for reliability assessment for FastICA.

Icasso estimates a large number of independent components with changing initial conditions and/or bootstrapping, and visualizes their clustering in the signal space. Each estimated independent component is one point in the space. If an independent component is stable, (almost) every run of the algorithm should produce a point that is very close to the ideal component corresponding to the cluster center. Reliable independent components correspond to tight clusters, and unreliable ones correspond to points which do not belong to any cluster. Preliminary results with a biomedical benchmarking data set are reported in [1]. See also Fig. 2.3.

References

[1] J. Himberg and A. Hyvärinen. (2003). Icasso: software for investigating the reliability of ICA estimates by clustering and visualization. In *Proc. Int. workshop of Neural Networks for Signal Processing (NNSP2003)*, Toulouse, France.

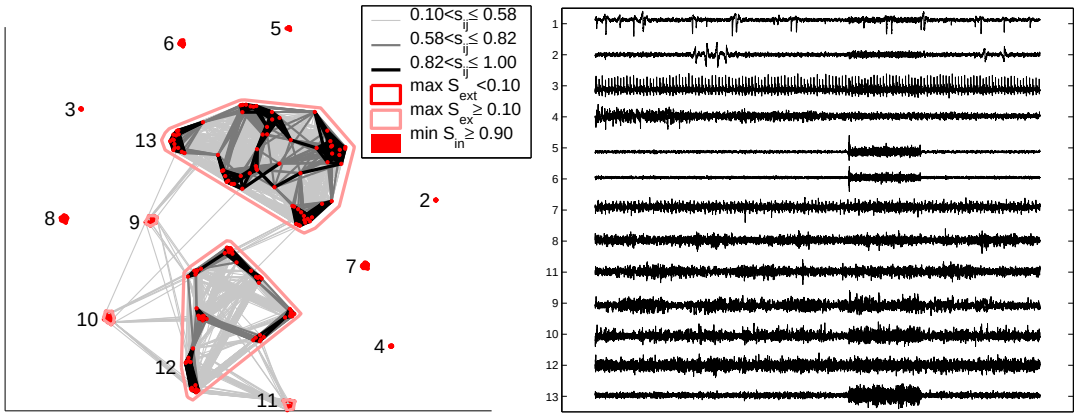
[2] A. Hyvärinen. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions of Neural Networks*, 10(3):626–635.

[3] F. Meinecke, A. Ziehe, M. Kawanabe, and K.-R. Müller (2002). A resampling approach to estimate the stability of one-dimensional or multidimensional independent components. *IEEE Transactions on Biomedical Engineering*, 49(12):1514–1525.

[4] R. Vigário, V. Jousimäki, M. Hämäläinen, R. Hari, and E. Oja. (1998). Independent component analysis for identification of artifacts in magnetoencephalographic recordings. In *Advances in Neural Information Processing Systems*, vol. 10, pp. 229–235. MIT Press.

[1] A popular public domain software package is available at `www.cis.hut.fi/research/software.shtml`

[2] at `www.cis.hut.fi/jhimberg/icasso`



(a) (b)

Figure 2.3: A MEG data set from [4] is analyzed with FastICA and Icasso. Panel (a) shows the correlations of all ICA estimates; 13 clusters have been selected and labeled #1–#13. One estimate from each is selected, the one that is nearest to the centroid of the cluster. The estimates are presented in panel (b) ordered according to the dispersion of the clusters. From the previous studies, we know that source estimates corresponding to clusters #1 and #2 correspond to eye movements, #3 to heart and #9 to the digital watch. Sources #5 and #6 are related to muscular activities due to biting. Source #4 is interesting since it is clearly well estimated but the physiological explanation is not yet known.

2.8 The European joint project BLISS

Juha Karhunen, Erkki Oja, Harri Valpola, Ricardo Vigario, Antti Honkela, Jaakko Särelä

Our laboratory has been one of the five participants in a large European joint project on Blind Source Separation and Applications, abbreviated BLISS. The project originally covered three years between June 2000 and June 2003, but it was later on extended by five months up to October 2003. The total funding of the BLISS project was 1.2 million euros, and it belonged to the "Information Society Technologies" programme (1998-2002) funded by the European Community. The other participating institutes and the leaders of the BLISS project there were:

- INESC, Lisbon, Portugal (Prof. Luis Almeida, coordinator);
- INPG (Inst. Nat. Polytechnique de France), Grenoble, France (Profs. Christian Jutten and Dinh-Tuan Pham);
- GMD First (Fraunhofer Institute), Berlin, Germany (Prof. Klaus-Robert Müller);
- McMaster University, Hamilton, Canada (Prof. Simon Haykin; adjunct member, getting its funding from Canada).

INESC withdrew from the project in summer 2002, and after that INPG has acted as a new coordinator. In addition, the project had both industrial and scientific advisory board members.

The project divided into two major parts: Theory and Algorithms, and Applications. The first part Theory and Algorithms consisted of three subprojects, which are Linear ICA, Nonlinear Separation, and Nonlinear BSS (Blind Source Separation). Our laboratory has formally been involved in the last two subprojects, but we contributed also to the subproject on linear ICA. The second major part Applications had two subprojects, Biomedical Applications and Acoustic Mixtures, and our laboratory participated in the first subproject on biomedical applications.

Meetings of the participants of the project have been arranged mainly in context with conferences and other events. The second official review meeting of the BLISS project was held in Brussels, Belgium, in July 2002. There it was decided to extend the project by five months for achieving its goals, and the workplan was revised appropriately. The final official review meeting was held in Paris at the end of October 2003. The reviewers were quite satisfied with the final results of the project, stating that all the major goals of the BLISS project have been achieved.

Especially during the later part of the project, practical co-operation between the participating laboratories has been deepened by researcher visits and writing of joint publications. Two major events intended to European researchers and graduate students have been organized largely within the BLISS project: the European Meeting on Independent Component Analysis in Vietri sul Mare, Italy, in February 2002; and the European Summer School on ICA in Berlin, Germany, in June 2003. Both events were quite successful with tens of interested participants.

The research carried out in our laboratory using the BLISS project funding is described in the chapters "Independent component analysis and blind source separation", "Variational Bayesian learning of generative models", and "Analysis of independent components in biomedical signals", as well as in the many associated publications. The results have been reported also in the deliverables and reports of the project. More information on the

BLISS project is available on its homepage [1] where one can find final reports and other deliverables, data sets, as well as software.

References

[1] Homepage of the BLISS project `http://www.lis.inpg.fr/pages_perso/bliss/`.

Chapter 3

Variational Bayesian learning of generative models

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

3.1 Bayesian modeling and variational learning

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

$$p(\boldsymbol{\theta}|X, \mathcal{H}) = \frac{p(X|\boldsymbol{\theta}, \mathcal{H})p(\boldsymbol{\theta}|\mathcal{H})}{p(X|\mathcal{H})} \tag{3.1}$$

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 solves optimally 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 (3.1) 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. We use a particular variational method known as ensemble learning [3, 4] that has recently become very popular because of its good properties. In ensemble learning, 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

$$D(q \parallel p) = \int q(v) \ln \frac{q(v)}{p(v)} dv \tag{3.2}$$

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

A key idea in ensemble learning 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 ensemble learning is that it avoids overfitting which would be a difficult problem if ML or MAP estimates were used. Ensemble learning 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.

Ensemble learning 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. Ensemble learning was originally derived from information theoretic point of view in [3]. The information theoretic view also provides insights to many aspects of learning and helps explain several common problems [5].

In the following subsections, we first present some recent theoretical improvements to ensemble learning methods and a practical building block framework that can be used to easily construct new models. After this we discuss practical models for nonlinear static and dynamic blind source separation as well as hierarchical modeling of variances. Finally we present applications of the developed Bayesian methods to inferring missing values from data and to detection of changes in process states.

References

[1] C. Bishop, *Neural Networks for Pattern Recognition*. Clarendon Press, 1995.

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

[3] G. Hinton and D. van Camp. Keeping neural networks simple by minimizing the description length of the weights. In *Proc. of the 6th Annual ACM Conf. on Computational Learning Theory*, pages 5–13, Santa Cruz, California, USA, 1993.

[4] D. MacKay. Developments in Probabilistic Modelling with Neural Networks – Ensemble Learning. In *Neural Networks: Artificial Intelligence and Industrial Applications.*

Proc. of the 3rd Annual Symposium on Neural Networks, pages 191–198, Nijmegen, Netherlands, 1995.

[5] A. Honkela and H. Valpola. Variational learning and bits-back coding: an information-theoretic view to Bayesian learning. *IEEE Transactions on Neural Networks*, 2004. To appear.

3.2 Theoretical improvements

Using pattern searches to speed up learning

The parameters of a latent variable model used in unsupervised learning can usually be divided into two sets: the latent variables or sources $\boldsymbol{S}$, and other model parameters $\boldsymbol{\theta}$. The learning algorithms used in variational Bayesian learning of these models have traditionally been variants of the expectation maximization (EM) algorithm, which is based on alternatively estimating $\boldsymbol{S}$ and $\boldsymbol{\theta}$ given the present estimate of the other.

The standard update algorithm can be very slow in case of low noise, because the updates needed for $\boldsymbol{S}$ and $\boldsymbol{\theta}$ are strongly correlated. Therefore one set can be changed very little only while the other is kept fixed in order to preserve the reconstruction of the data. So-called pattern searches, which use the combined direction of a round of standard updates and then perform a line search in this direction, can help to avoid this problem [1].

The effect of using pattern searches is demonstrated in Figure 3.1 which shows the speedups attained in experiments with hierarchical nonlinear factor analysis (HNFA) (see Sec. 3.4) in different phases of learning. As the nonlinear model is susceptible to local minima, the different algorithms do not always converge to the same point. So the comparison was made by looking at the times required by the methods to reach a certain level of the cost function value above the worst local minimum found by the two algorithms.

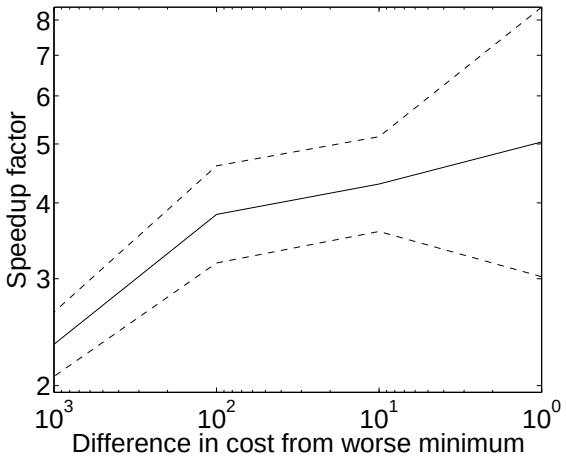


Figure 3.1: The average speedup obtained by pattern searches in different phases of learning. The speedup is measured by the ratio of times required by the basic algorithm and pattern search method to reach certain level of cost function value. The solid line shows the mean of the speedups over 20 simulation with different initializations and the dashed lines show 99 % confidence intervals for the mean.

Effect of posterior approximation

Most applications of ensemble 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 to represent all different solutions, which may greatly affect the found solution.

Our recent paper [2] shows that neglecting the posterior correlations of the sources in $q(\boldsymbol{S})$ introduces a bias in favor of 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.

Figure 3.2 illustrates this 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. Therefore, the cost of inaccurate assumption would increase towards the ICA solution as shown with the dashed line on the second plot of Fig. 3.2. On the other hand, if the true mixing matrix is not orthogonal, the optimal posterior covariance of the sources could look like the one in the upper plot of Fig. 3.2. Then, the misfit of the posterior approximation of the sources is minimized in the PCA solution where the true posterior covariance would be diagonal.

In [2], we considered a linear dynamic ICA model but the analysis extends to nonlinear mixtures and non-Gaussian source models as well.

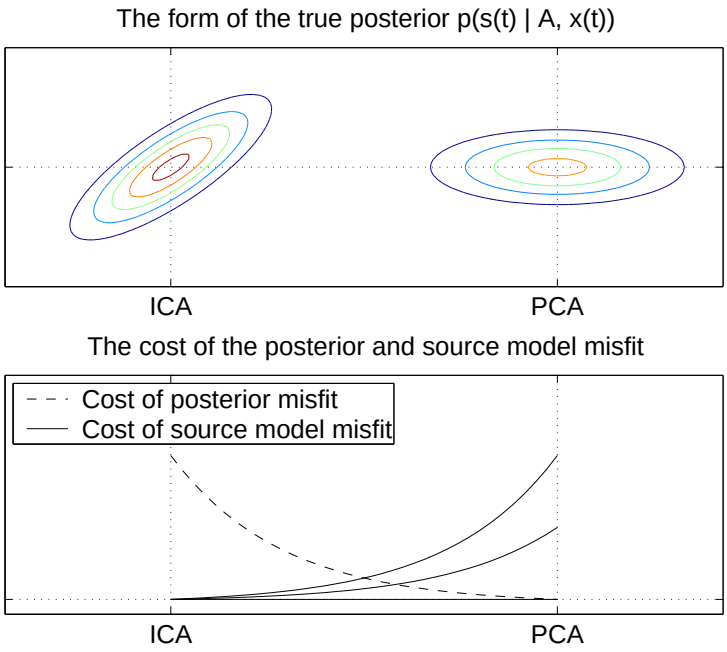


Figure 3.2: Schematic illustration of the trade-offs between the ICA and PCA solutions. In the PCA solution, the posterior covariance of the sources is diagonal. This minimizes the misfit between the optimal posterior and its approximation. However, the sources are explained better in the ICA solution.

References

[1] A. Honkela, H. Valpola, and J. Karhunen. Accelerating cyclic update algorithms for parameter estimation by pattern searches. *Neural Processing Letters*, 17(2):191–203, 2003.

[2] A. Ilin, and H. Valpola. On the Effect of the Form of the Posterior Approximation in Variational Learning of ICA Models. In *Proc. of the 4th Int. Symp. on Independent Component Analysis and Blind Signal Separation (ICA2003)*, pages 915–920, Nara, Japan, 2003.

3.3 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 [1]. The theoretical background on which it is based on, was published in [2].

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. Each of them 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.

Figure 3.3 shows an example of a structure which can be built using the Bayes Blocks library. More structures can be found in [2, 3], and their application to hierarchical modeling of variances is described in Sec. 3.5.

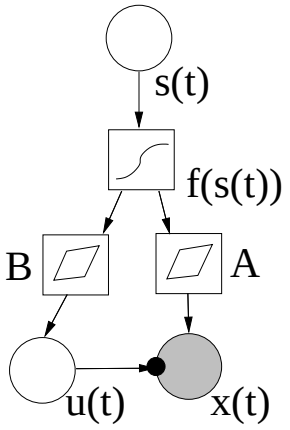


Figure 3.3: An example of a structure that can be built using Bayes Blocks. It includes latent variables $s(t)$, $u(t)$ (and some parameters), observed variables $x(t)$, a nonlinearity $f(\cdot)$, and affine transformations A and B. The variables $u(t)$ model the variances of $x(t)$.

References

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

[2] 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.

[3] Harri Valpola, Markus Harva, and Juha Karhunen. Hierarchical models of variance sources. *Signal Processing*, 84(2):267–282, 2004.

3.4 Nonlinear static and dynamic blind source separation

The linear principal and independent component analysis (PCA and ICA, respectively) [1] model the data so that it has been generated by sources through a linear mapping. PCA looks for uncorrelated sources, restricting the directions of the sources to be mutually orthogonal. On the other hand, ICA requires that the sources are statistically independent which is a stronger assumption than uncorrelatedness, but there is no orthogonality restriction. In general, PCA is sufficient for Gaussian sources only, because it does not exploit higher than second-order statistics in any way [1].

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

$$\mathbf{x}(t) = \mathbf{f}(\mathbf{s}(t), \boldsymbol{\theta}_f) + \mathbf{n}(t)\,. \tag{3.3}$$

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 $\boldsymbol{\theta}_f$.

Nonlinear ICA by multi-layer perceptrons

In an earlier work [2, 3] we have used multi-layer perceptron (MLP) network with tanh-nonlinearities to model the mapping $\mathbf{f}$:

$$\mathbf{f}(\mathbf{s}; \mathbf{A}, \mathbf{B}, \mathbf{a}, \mathbf{b}) = \mathbf{B}\tanh(\mathbf{A}\mathbf{s} + \mathbf{a}) + \mathbf{b}\,. \tag{3.4}$$

The mapping $\mathbf{f}$ is thus parametrized by the matrices $\mathbf{A}$ and $\mathbf{B}$ and bias vectors $\mathbf{a}$ and $\mathbf{b}$. MLP networks are well suited for nonlinear PCA and ICA. 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, close to linear mappings with them. This makes it possible to learn high dimensional nonlinear representations in practice.

Traditionally MLP networks have been used for supervised learning where both the inputs and the desired outputs are known. Here sources correspond to inputs and observations correspond to desired outputs. The sources are unknown and therefore learning is unsupervised.

Usually the linear PCA and ICA models do not have an explicit noise term $\mathbf{n}(t)$ and the model is thus simply $\mathbf{x}(t) = \mathbf{f}(\mathbf{s}(t)) = \mathbf{A}\mathbf{s}(t) + \mathbf{a}$, where $\mathbf{A}$ is a mixing matrix and $\mathbf{a}$ is a bias vector. The corresponding PCA and ICA models which include the noise term are often called factor analysis and independent factor analysis (FA and IFA) models. The nonlinear models discussed here can therefore also be called nonlinear factor analysis and nonlinear independent factor analysis models.

Hierarchical nonlinear factor analysis

The computational complexity of the variational Bayesian learning algorithm for the MLP network model is quadratic with respect to the number of sources in the model. To avoid this problem, an alternative hierarchical structure based on the build block approach presented in Section 3.3 was studied in [4]. One of the building blocks is a Gaussian variable ξ followed by a nonlinearity ϕ:

$$\phi(\xi) = \exp(-\xi^2)\,. \tag{3.5}$$

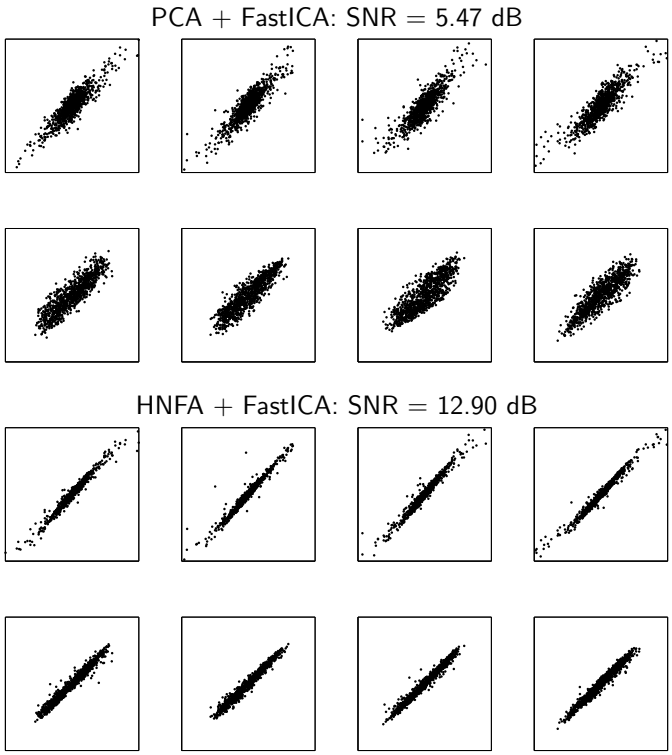


Figure 3.4: Each scatter plot shows the values of one original source signal plotted against the best corresponding estimated source signal after a rotation with FastICA.

The motivation for choosing this particular nonlinearity is that for Gaussian posterior approximation $q_\xi(\xi)$, the posterior mean and variance and consequently the cost function can be evaluated analytically.

Using this construction—Gaussian variables followed by nonlinearity—it is possible to build nonlinear mappings for which the learning time is linear with respect to the size of the model. The key idea is to introduce latent variables $\mathbf{h}(t)$ before the nonlinearities and thus split the mapping Eq. (3.4) into two parts in the hierarchical nonlinear factor analysis (HNFA) model:

$$\mathbf{h}(t) = \mathbf{B}\mathbf{s}(t) + \mathbf{b} + \mathbf{n}_h(t) \tag{3.6}$$

$$\mathbf{x}(t) = \mathbf{A}\phi[\mathbf{h}(t)] + \mathbf{C}\mathbf{s}(t) + \mathbf{a} + \mathbf{n}_x(t)\,, \tag{3.7}$$

where $\mathbf{n}_h(t)$ and $\mathbf{n}_x(t)$ are Gaussian noise terms. Note that we have included a short-cut mapping $\mathbf{C}$ from sources to observations. This means that hidden nodes only need to model the deviations from linearity.

The source model in HNFA is Gaussian so the model cannot be used for ICA. It can, however, be used as a nonlinear preprocessing method that extracts the correct subspace within which the correct rotation is then recovered using a standard linear ICA algorithm. Figure 3.4 illustrates the results of using HNFA together with the FastICA algorithm [1] for standard linear ICA to extract the sources from an artificial noisy nonlinear mixture. The data set used consisted of 1000 20-dimensional vectors which were created by nonlinearly mixing eight non-Gaussian independent random sources. A nonlinear model is clearly required in order to capture to underlying nonlinear manifold.

Nonlinear state-space models

In many cases, measurements originate from a dynamic 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 $\mathbf{s}(t)$ [5]. 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. (3.3) can easily be extended to a dynamic one by adding another nonlinear mapping to model the dynamics. This leads to source model

$$\mathbf{s}(t) = \mathbf{g}(\mathbf{s}(t-1), \boldsymbol{\theta}_g) + \mathbf{m}(t)\,, \tag{3.8}$$

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

As in nonlinear factor analysis, the nonlinear functions are modeled by MLP networks. The mapping $\mathbf{f}$ has the same functional form (3.4). 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:

$$\mathbf{g}(\mathbf{s}(\mathbf{t}-\mathbf{1})) = \mathbf{s}(t-1) + \mathbf{D}\tanh[\mathbf{C}\mathbf{s}(t-1) + \mathbf{c}] + \mathbf{d}\,. \tag{3.9}$$

An important advantage of the proposed new method is its ability to learn a high-dimensional latent source space. We have also reasonably solved computational and overfitting 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. A process monitoring application is discussed in Section 3.6 in more detail.

Postnonlinear factor analysis

Our recent work restricts the general nonlinear mapping in (3.3) to the important case of post-nonlinear (PNL) mixtures. The PNL model consists of a linear mixture followed by componentwise nonlinearities acting on each output independently from the others:

$$x_i(t) = f_i\left[\mathbf{A}_{i,:}\mathbf{s}(t)\right] + n_i(t) \qquad i = 1, \ldots, n \tag{3.10}$$

The notation $\mathbf{A}_{i,:}$ in this equation means the i:th row of the mixing matrix $\mathbf{A}$. Preliminary results from the model (3.10) are encouraging. The results will be published in forthcoming papers.

References

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

[2] Harri Lappalainen and Antti 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.

[3] H. Valpola, E. Oja, A. Ilin, A. Honkela, and J. Karhunen. Nonlinear blind source separation by variational Bayesian learning. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, E86-A(3):532–541, 2003.

[4] 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.

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

3.5 Hierarchical modeling of variances

In many models, variances are assumed to be constant 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 using sampling the difficulties with infinite probability densities are avoided, but these methods are not efficient enough for very large datasets. Our variational Bayesian method [1, 2], which is based on our building blocks framework (see Sec. 3.3), is able to jointly model both variances and means efficiently.

The basic building block in our models is the variance neuron, 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 3.5 shows three examples of usage of variance neurons. The first model does not have any upper layer model for the variances. There the variance neurons are useful as such for generating super-Gaussian distributions for $\mathbf{s}$, enabling in effect 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 neurons $\mathbf{u}_x$. 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 $\mathbf{r}(t)$.

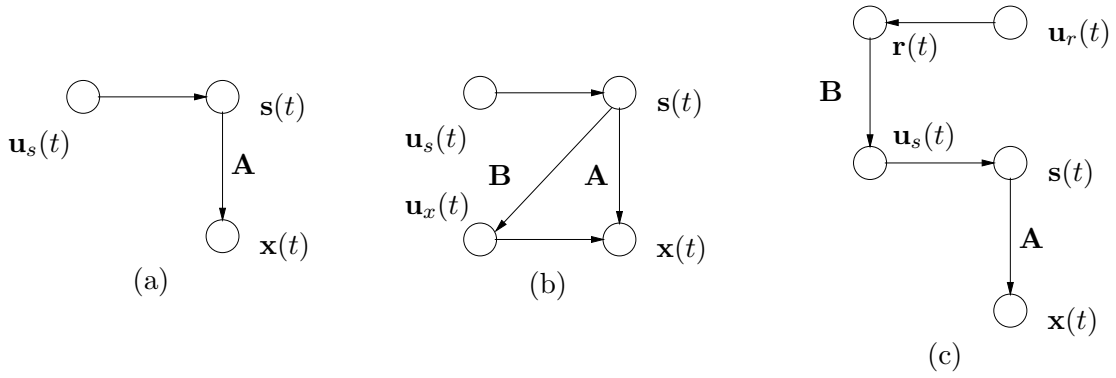


Figure 3.5: Various model structures utilizing variance neurons. Observations are denoted by $\mathbf{x}$, linear mappings by $\mathbf{A}$ and $\mathbf{B}$, sources by $\mathbf{s}$ and $\mathbf{r}$, and variance neurons by $\mathbf{u}$.

We have used the model of Fig. 3.5(c) for finding variance sources from biomedical data containing MEG measurements from a human brain. Part of that dataset is shown in Figure 3.6(a). 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.

Some of the estimated ordinary sources $\mathbf{s}(t)$ and their variance neurons $\mathbf{u}_s(t)$ are shown in Figures 3.6(b) and 3.6(c). The variance sources $\mathbf{r}(t)$ that were discovered are shown in Figure 3.6(d). 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 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.

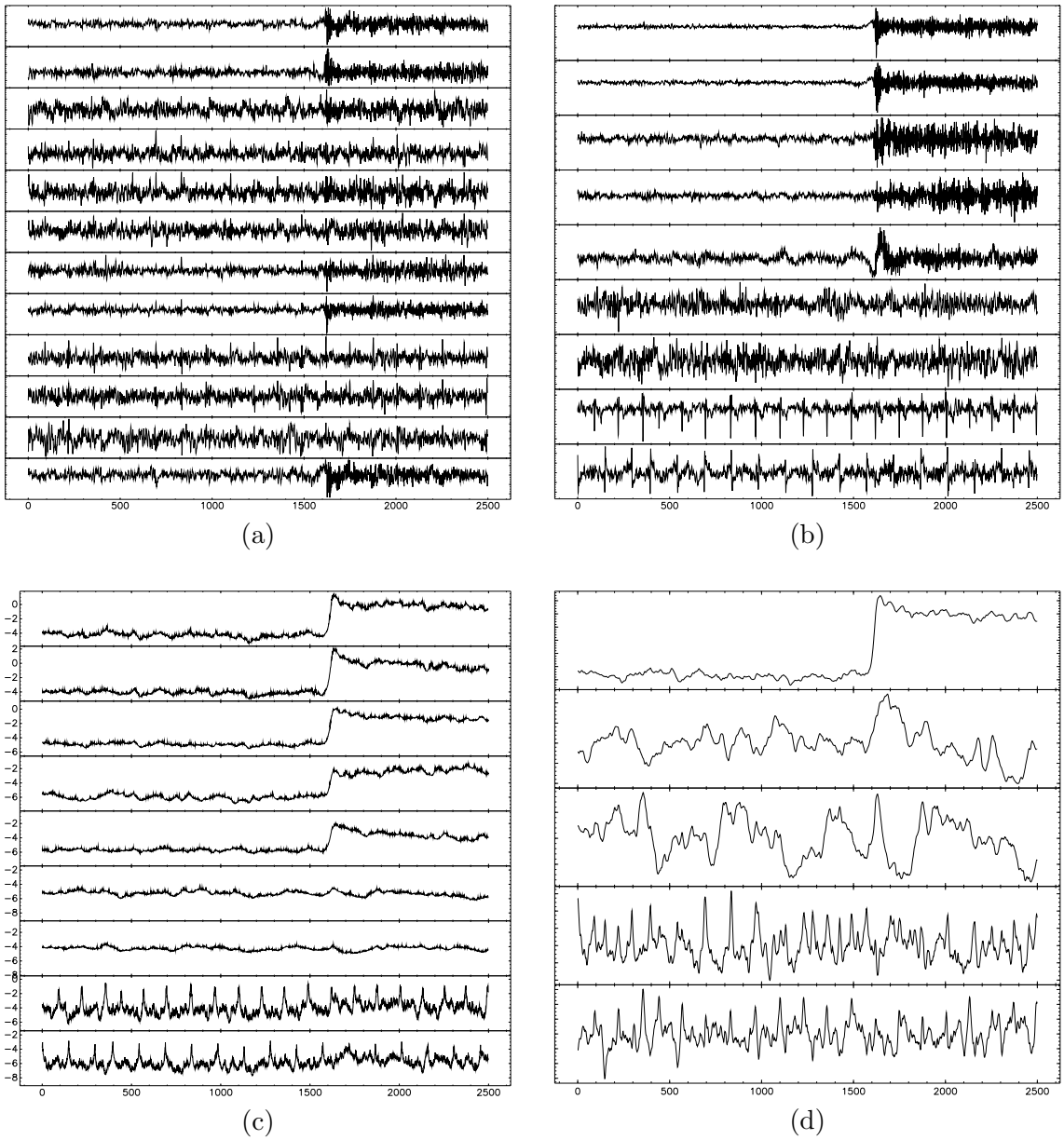


Figure 3.6: (a) MEG recordings (12 out of 122 time series). (b) Sources $\mathbf{s}(t)$ (nine out of 50) estimated from the data. (c) Variance neurons $\mathbf{u}_s(t)$ corresponding to the sources. (d) Variance sources $\mathbf{r}(t)$ which model the regularities found from the variance neurons.

References

[1] Harri Valpola, Markus Harva, and Juha Karhunen. Hierarchical models of variance sources. *Signal Processing*, 84(2):267–282, 2004.

[2] Harri Valpola, Markus Harva, and Juha Karhunen. Hierarchical models of variance sources. In *Proc. 4th Int. Symp. on Independent Component Analysis and Blind Signal Separation (ICA2003)*, pages 83–88, Nara, Japan, 2003.

3.6 Applications

In this section, applications of hierarchical nonlinear factor analysis and nonlinear state-space models discussed earlier in Section 3.4 are presented.

Missing values

Generative models can usually easily deal with missing observations. For instance in self-organizing maps (SOM) the winning neuron can be found based on those observations that are available. The generative model can also be used to fill in the missing values. This way unsupervised learning can be used for a similar task as supervised learning. Both the inputs and desired outputs of the learning data are treated equally. When a generative model for the combined data is learned, it can be used to reconstruct the missing outputs for the test data. The scheme used in unsupervised learning is more flexible because any part of the data can act as the cue which is used to complete the rest of the data. In supervised learning, the inputs always act as the cue.

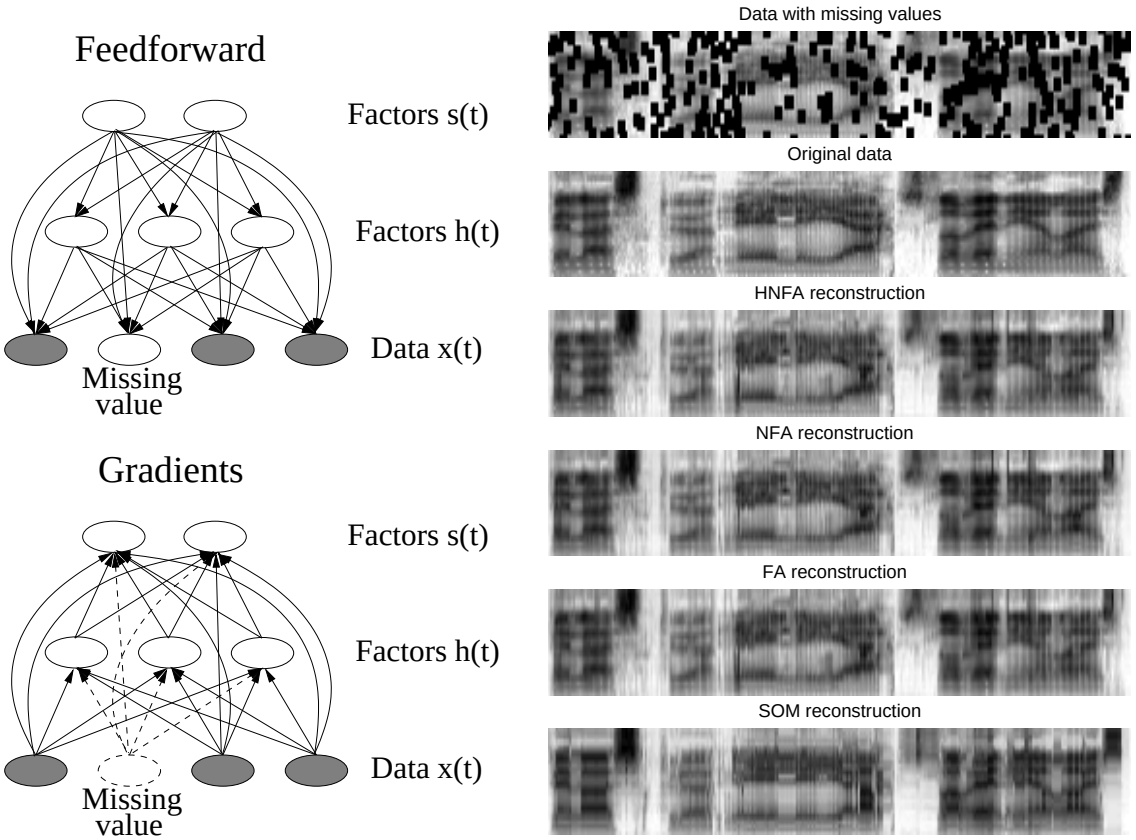


Figure 3.7: Left: The reconstructions of missing values using HNFA are produced in the feedforward direction. In the gradient direction, the components with missing values do not affect the factors, which are thus inferred using only the observed data. Right: Speech data reconstruction example with best parameters of each algorithm.

The quality of the reconstructions provides insight to the properties of different unsupervised models. The ability of self-organizing maps, linear principal component analysis, nonlinear factor analysis, and hierarchical nonlinear factor analysis to reconstruct the missing values of various data sets have been studied in [1]. Experiments were conducted

Missing value pattern	FA	HNFA	NFA	SOM
patches	1.87	1.80 ± 0.03	1.74 ± 0.02	1.69 ± 0.02
patches, permutated	1.85	1.78 ± 0.03	1.71 ± 0.01	1.55 ± 0.01
randomly	0.57	$0.55 \pm .005$	$0.56 \pm .002$	0.86 ± 0.01
randomly, permutated	0.58	$0.55 \pm .008$	$0.58 \pm .004$	0.87 ± 0.01

Table 3.1: Mean-square reconstruction errors and their standard deviations for speech spectra containing various types of missing values.

using four different patterns for the missing values. This way, different aspects of the algorithms could be studied.

Table 3.1 shows the mean-square reconstruction errors (and their standard deviations over different runs) of missing values in speech spectra. In Figure 3.7, the reconstructed spectra are shown for a case where data were missing in patches and the data is not permutated. In the permutated case, the learning data contained samples which were similar to the test data with missing values. This task does not require generalization but rather memorization of the learned data. SOM performs the best in this task because it has the largest amount of parameters in the model.

The task where the data was missing randomly and not in patches of several neighboring frequencies does not require a very nonlinear model but rather an accurate representation of a high-dimensional latent space. Linear and nonlinear factor analysis perform better than SOM whose parametrization is not well suited for very high-dimensional latent spaces. The conclusion of these experiments was that in many respects the properties of (hierarchical) nonlinear factor analysis are closer to linear factor analysis than highly nonlinear mappings such as SOM. The nonlinear extensions of linear factor analysis are nevertheless able to capture nonlinear structure in the data and perform as well or better than linear factor analysis in all the reconstruction tasks. The new hierarchical nonlinear factor analysis did not outperform the older nonlinear factor analysis in reconstruction accuracy, but it was more reliable and computationally lighter.

Detection of process state changes

One potential application for the nonlinear dynamic state-space model discussed in Section 3.4 is process monitoring. In [2, 3], ensemble 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 3.8 and 3.9 show an experiment with a change generated in the middle of the new data set, at time 1500, when the underlying dynamics of one of the Lorenz processes abruptly changes. Even though it is very difficult to detect this from the observed nonlinear mixtures shown in Fig. 3.8, the change detection method based on the estimated model readily detects the change raising alarms after the time of change. The method is also able to find out in which states the change occurred: Analyzing the structure of the cost function helps in localizing the detected changes, as demonstrated in Fig. 3.9.

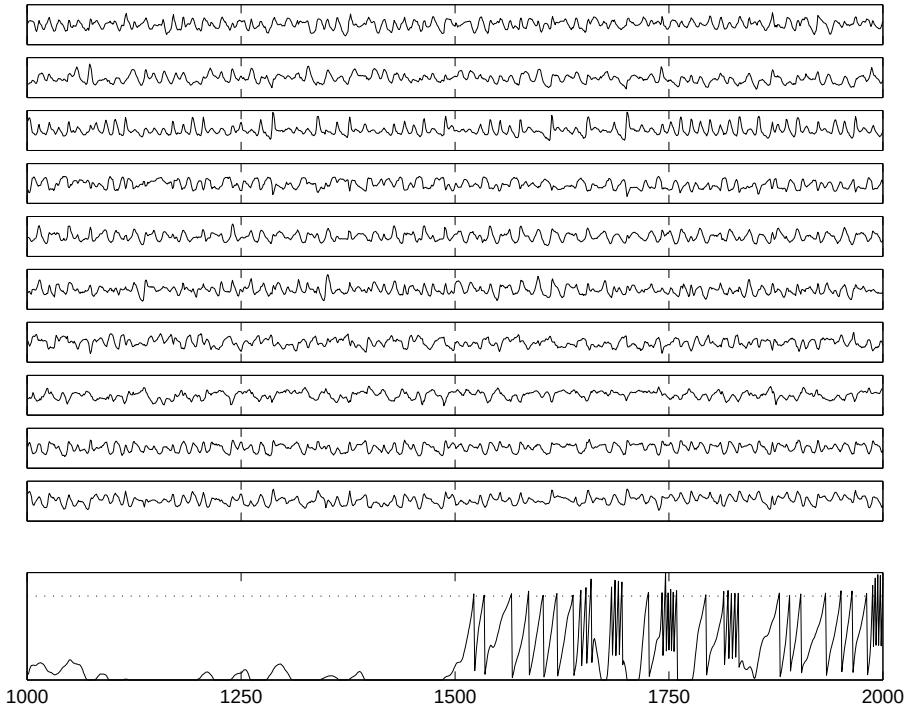


Figure 3.8: The monitored process (10 time series above) with the change simulated at $t = 500$. Even though the change is hardly visible to the eye, it has been detected with the estimated model: The test statistic of the proposed method is shown below.

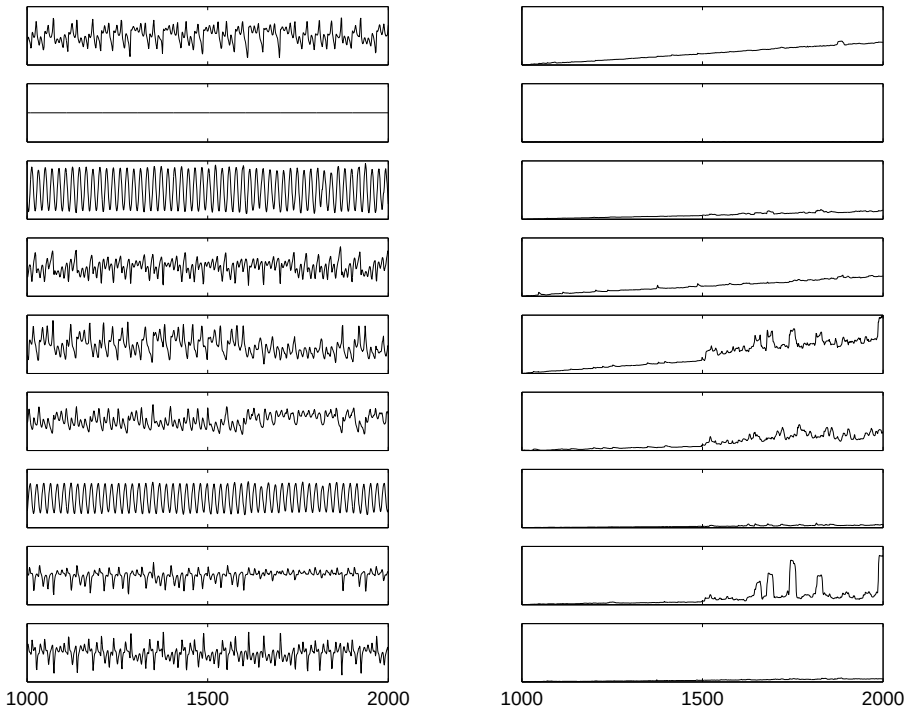


Figure 3.9: The estimated process states (left) and their contribution to the cost function (right). The cause of the change can be verified by detecting the states which increase their cost contribution.

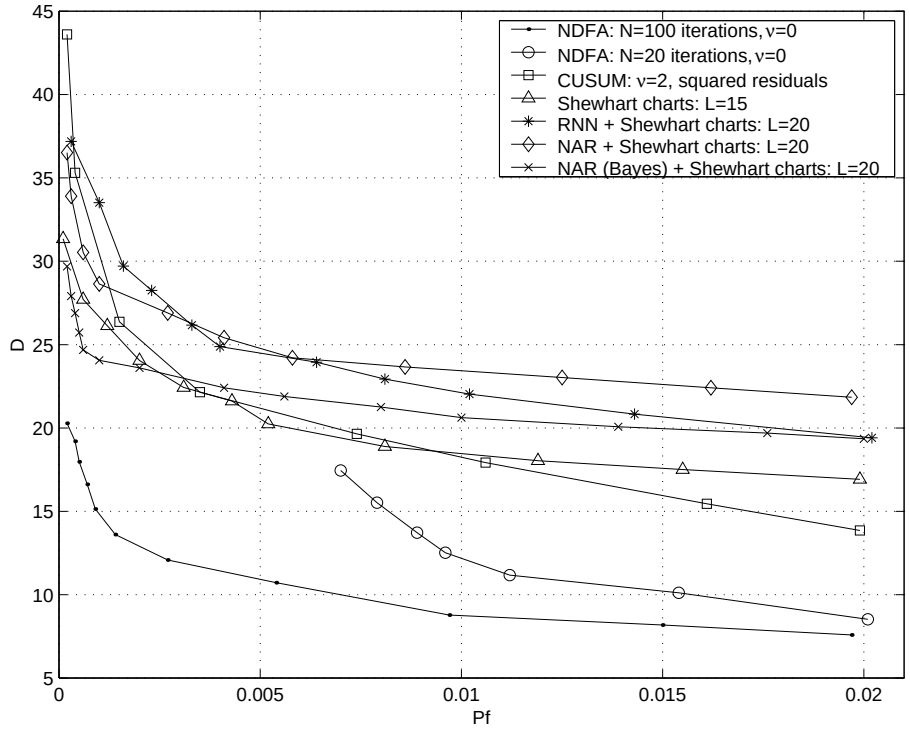


Figure 3.10: Two change detection performance measures calculated for the proposed NDFA method as well as for some alternative techniques. The probability of false alarms P_f is plotted against the average time to detection D for different values of the detection threshold. The closer a curve is to the origin, the faster the algorithm can detect the change with low false alarm rate.

The experimental results in Fig. 3.10 show that the method outperforms several other change detection techniques. Two alternative approaches compared with our method are based on other types of nonlinear dynamic models, namely nonlinear autoregressive (NAR) model and recurrent neural network (RNN) model. Other compared methods monitored simple indicators of the process such as its mean and covariance matrix (CUSUM algorithm and Shewhart control charts).

References

[1] T. Raiko, H. Valpola, T. Östman and J. Karhunen. Missing values in hierarchical nonlinear factor analysis. In *Proc. of the Int. Conf. on Artificial Neural Networks and Neural Information Processing - ICANN/ICONIP 2003*, pages 185–189, Istanbul, Turkey, June 2003.

[2] A. Iline, H. Valpola, and E. Oja. Detecting process state changes by nonlinear blind source separation. In *Proc. of the 3rd Int. Workshop on Independent Component Analysis and Signal Separation (ICA2001)*, pages 704–709, San Diego, California, December 2001.

[3] A. Ilin, H. Valpola, and E. Oja. Nonlinear Dynamical Factor Analysis for State Change Detection. *IEEE Transaction on Neural Networks*, 2004. To appear.

Chapter 4

Computational neuroscience

Aapo Hyvärinen, Patrik Hoyer, Jarmo Hurri, Mika Inki

4.1 The statistical structure of natural images and visual representation

Our research concentrated on modelling visual perception using statistical models. This work was a direct continuation of what was reported in the previous biennial report.

Our basic approach is to build models of the statistical structure of the typical input that the perceptual system receives, and estimate the parameters of the model from realistic input, such as digital photographs of wild-life scenes, or digital video. We then describe the function of parts of the visual cortex as statistical estimation and inference in such models. Our modelling has been largely based on extensions of independent component analysis and blind source separation methods. The goal is typically to transform a data vector into components that are statistically independent or whose dependency structure is quite simple.

We have developed a number of new models that extend the now well-known results on independent components of natural images:

Non-negative sparse coding This model is specialized to analysis of data that is non-negative. More precisely, the data is usually positive and close to zero, and occasionally gets large positive values [1,2]. The model seems especially suited for analysis of higher-order features that are computed from outputs of lower non-negative features, such as complex cells.

Models of variance dynamics It is well-known that the independent components of natural images are not independent. Our previous work already modelled some of the dependencies that remain after ICA. Yet, no existing model have been able to estimate a full two-layer model of natural images, where the seoncd layer explains some of the dependencies left after the first linear layer. Previous research has only been able to estimate two-layer models when one of the layers has been fixed. We developed a model where two layers can be estimated based on the temporal structure of natural image sequences [4]. The layers correspond to simple and complex cells in the primary visual cortex. See Fig. 4.1 for an illustration of the main kinds of dependencies, and Fig. 4.2 for some dependencies estimated from real data.

Bubble coding We have proposed a unifying framework [6] for several models of the statistical structure of natural image sequences. The framework combines three properties: sparseness, temporal coherence, and energy correlations. It leads to models where the joint activation of the linear filters (simple cells) takes the form of "bubbles", which are regions of activity that are localized both in time and in space, space meaning the cortical surface or a grid on which the features are arranged. The concept of bubbles is closely related to invariant features such as those coded by complex cells; the principle is illustrated in Fig. 4.3.

Double-blind source separation These theoretical developments in biological modelling lead to the development of a new method of blind source separation [5]. The new method separates sources without the need for an explicit parametric model of their dependency structure. This is possible by some general assumptions on the structure of the dependencies: the sources are dependent only through their variances (general activity levels), and the variances of the sources have temporal correlations. The method can be called double-blind because of this additional blind aspect: We do not need to estimate

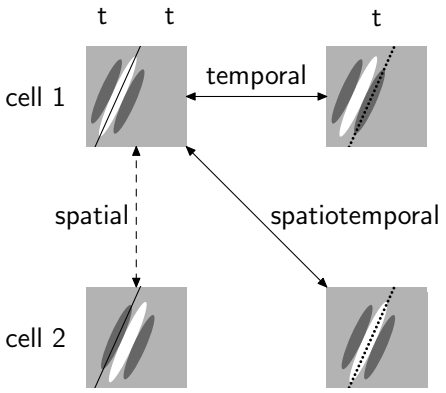


Figure 4.1: Illustration of the different types of dependencies found in natural image sequences. Consider a stimulus that consists of a line segment that moves accross the receptive fields of two linear neurons with receptive fields that have simile location, orientation and frequency. The outputs of a given neuron in two consecutive time steps are dependent. Further, two neurons with similar receptive fields have dependent outputs. Also, the outputs of similar neurons in consecutive time points are dependent.

Figure 4.2: Our two-layer model in [4] was able to estimate linear features and dependencies between the features. Each row shows the features with the highest and the lowest dependencies with respect to the reference feature on the left. Features with high dependencies code for similar orientation and frequeny. Features with low dependencies have very dissimilar parameters.

(or assume) a parametric model of the dependencies, which is in stark contrast to most previous methods.

Conditional and comparative statistics We have further investigated how the statistics are modified by conditioning by the value of one independent component [7]. If the components were really independent, this conditioning should not change anything, but our results show that it does. Comparison of the statistics of natural images with others kinds of images have also been performed [8].

Bayesian inference in the visual system On a more theoretical note, we proposed a model that explains some aspects of the response variability of neurons using the framework of Bayesian inference. It is proposed that the variability reflects a Monte Carlo sampling of the posterior probability distribution of perceptual parameters, given the input stimulus [3].

References

[1] P. O. Hoyer. Non-negative sparse coding. In *Proc. IEEE Workshop on Neural Networks for Signal Processing*, pages 557–565, 2002.

[2] P. O. Hoyer. Modeling receptive fields with non-negative sparse coding. In *Computational Neuroscience: Trends in Research 2003 (Proc. CNS 2002)*, 2003.

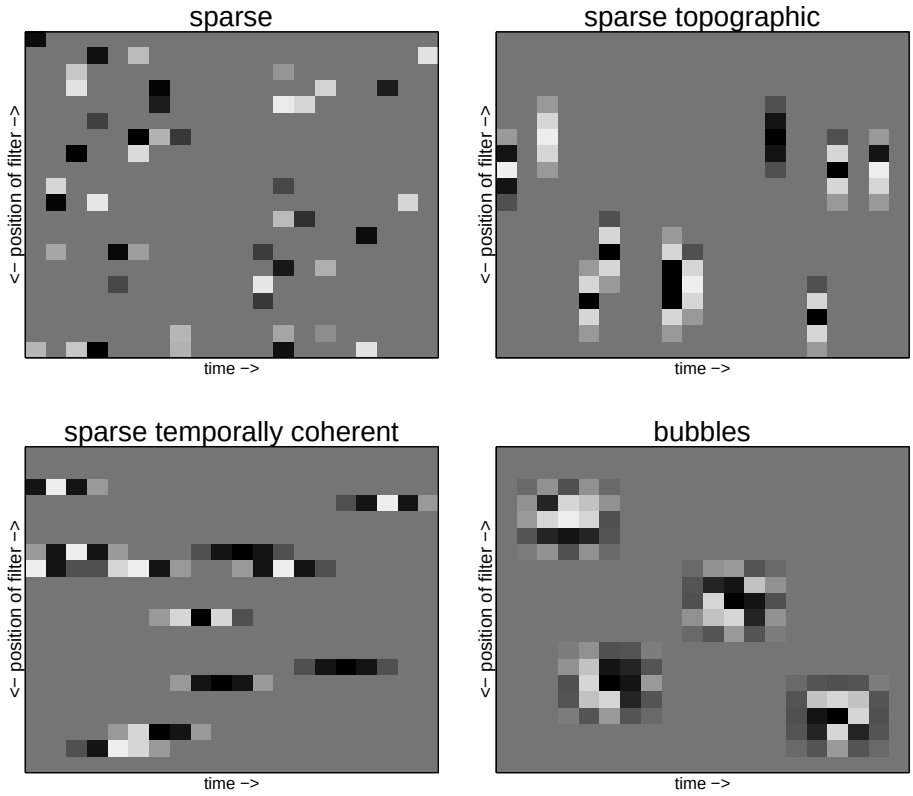


Figure 4.3: An illustration of a bubble representation. The plots show the outputs of filters as a function of time (horizontal axis) and the position of the filter on the topographic grid (vertical axis). Each pixel is the output of one unit at a given time point, gray being zero, white and black meaning positive and negative outputs. For simplicity, the topography is here one-dimensional. In the basic sparse representation, the filters are independent. In the topographic representation, the activations of the filters are also spatially grouped. In the representation that has temporal coherence, they are temporally grouped. The bubble representation combines all these aspects, leading to spatiotemporal activity bubbles. Note that the two latter types of representation require that the data has a temporal structure, unlike the two former ones.

[3] P. O. Hoyer and A. Hyvärinen. Interpreting neural response variablity as Monte Carlo sampling of the posterior. In *Advances in Neural Information Processing Systems*, volume 15, pages 277–284. MIT Press, 2003.

[4] J. Hurri and A. Hyvärinen. Temporal and spatiotemporal coherence in simple-cell responses: A generative model of natural image sequences. *Network: Computation in Neural Systems*, 14(3):527–551, 2003.

[5] A. Hyvärinen and J. Hurri. Blind separation of sources that have spatiotemporal variance dependencies. *Signal Processing*. In press.

[6] A. Hyvärinen, J. Hurri, and J. Väyrynen. Bubbles: A unifying framework for low-level statistical properties of natural image sequences. *J. of the Optical Society of America A*, 20(7):1237–1252, 2003.

[7] M. Inki. Examining the dependencies between ica features of image data. In *Proc. ICANN/ICONIP2003*, Istanbul, Turkey, 2003.

[8] M. Inki. ICA features of image data in one, two and three dimensions. In *Proc. Int. Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003)*, pages 861–866, Nara, Japan, 2003.

Chapter 5

Analysis of independent components in biomedical signals

Ricardo Vigário, Jaakko Särelä, Elina Karp, Jarkko Ylipaavalniemi

5.1 Biomedical data analysis

Ricardo Vigário, Jaakko Särelä, Elina Karp, Jarkko Ylipaavalniemi

In the period spanned by this report, we have enlarged our ongoing collaboration with the Brain Research Unit of the Helsinki University of Technology to the recently established Advanced Magnetic Imaging Centre, from the same university. We further explored contacts with the radiology department of Helsinki University Central Hospital. International collaborations were as well pursued with very positive outcomes.

The global list of publications, at the end of this report, contains further references to this work. The ones in this section should give a good starting point to the understanding of the results achieved within the project.

Coherence studies

Strong coherence around 20 Hz is known to exist between the MEG recording the primary motor cortex and the contralateral electromyogram (EMG) during isometric muscle contraction. In [2, 1], we applied a temporal decorrelation technique to identify the underlying brain areas producing signals coherent with the EMG. The algorithm chosen, the temporal decorrelation source separation (TDSEP [3]), exploits efficiently the temporal structure present in the data.

To reduce the occurrence of overlearning of the TDSEP algorithm in such high-dimensional data set (204 channels were recorded), a suitable reduction was performed during its whitening stage. Yet, we are now not necessarily interested in reconstructing the original data, but rather in the preservation of the corticomuscle coherence between the components and the reference EMG. Hence, we choose to sort the principal components according to their coherence power, discarding those that contribute the least to the coherence, in the vicinity of the characteristic 20 Hz.

Figure 5.1 presents the top 7 coherences between the 204 MEG channels and the right-hand EMG. The strongest coherence peaks around 14 Hz, which is close to the values suggested in the literature.

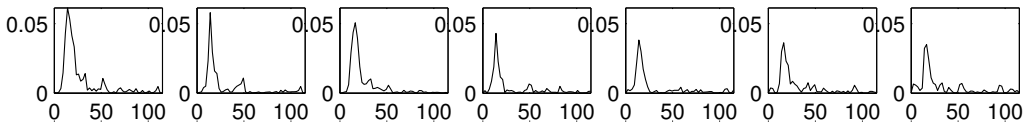


Figure 5.1: Strongest coherences between the MEG channels and the left-hand EMG.

In Fig. 5.2, the 7 most coherent TDSEP components are shown. Only the first signal shows significant coherence values (even higher than that of any single MEG channel). The

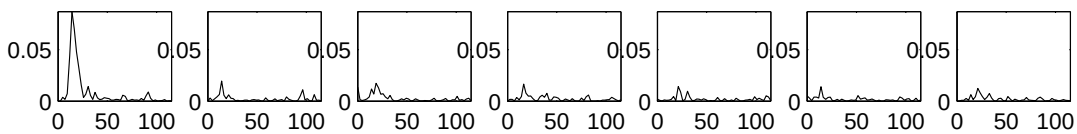
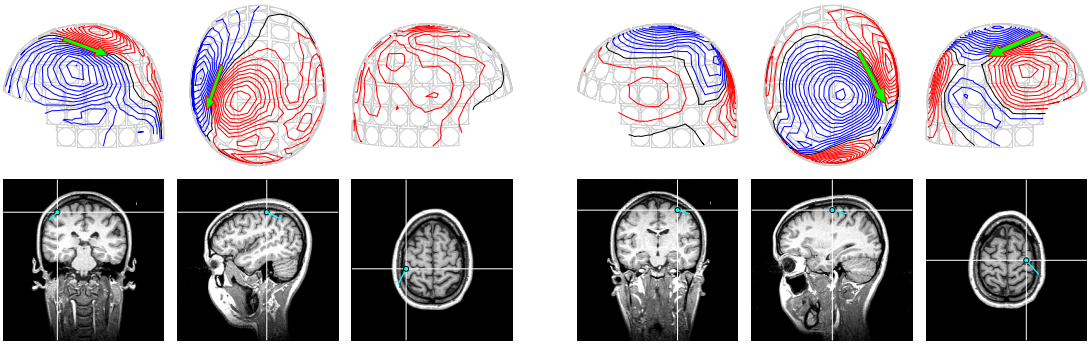


Figure 5.2: Coherence plots for the 7 most coherent TDSEP components.

field patterns associated with several TDSEP components, see Fig. 5.3, show selectivity over the central left and right hemispheres, and are well modeled by current dipoles in the respective primary motor cortices.



(a) (b)

Figure 5.3: Field patterns and equivalent current dipoles for the 1st TDSEP component (a) and the 25th (b).

References

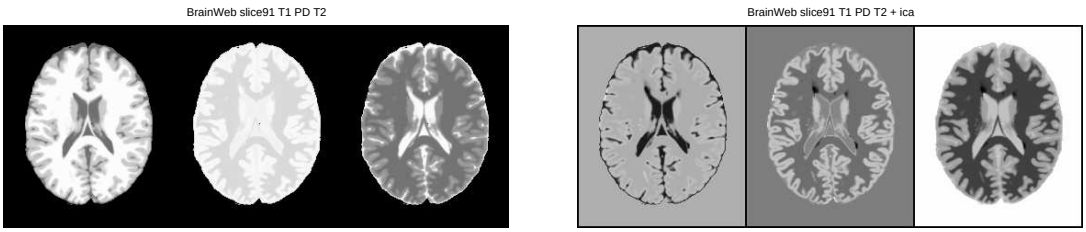
[1] Ricardo Vigário and Ole Jensen. Identifying cortical sources of corticomuscle coherence during bimanual muscle contraction by temporal decorrelation. In *Proc. of the 7th Int. Symp. on Signal Proc. and its Applications (ISSPA'03)*, pages 109 – 112, Paris, 2003.

[2] Ricardo Vigário, Ole Jensen, and Riitta Hari. Identifying cortical sources of corticomuscle coherence during bimanual muslce contraction by ICA. In *Proc. of the 13th Int. Conf. on Biomagnetism (BIOMAG'02)*, page 1005, Jena, Germany, 2002.

[3] Andreas Ziehe and Klaus-Robert Müller. TDSEP – an effective algrithm for blind separation using time structure. In *Proc. Int. Conf. on Artificial Neural Networks (ICANN'98)*, pages 675 – 680, Skövde, Sweden, 1998.

Structural MRI

Magnetic resonance imaging (MRI) is a non-invasive technique capable of producing 3D high resolution images of the human body. It relies on the interaction between the intrinsic magnetic fields of nuclei and strong externally applied magnetic fields and radiofrequency pulses. By adjusting internal parameters, related to the onset of the external pulses and fields, one can change the focus of the image on different anatomical and physiological properties of the tissues. A realistic simulation of those differences is depicted in Fig. 5.4a [1], for the standard T1, T2 and PD (proton density) parameter set. From the



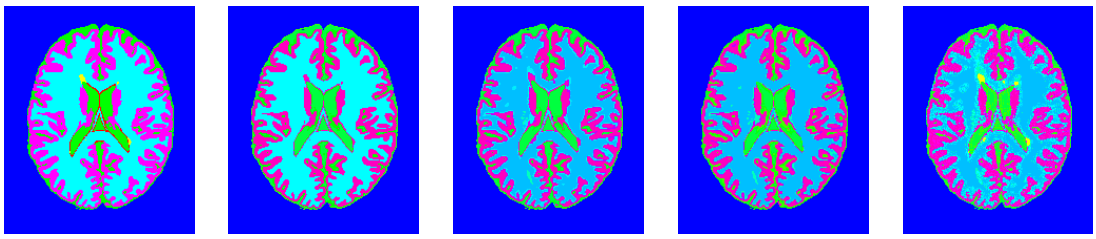
(a) (b)

Figure 5.4: Simulated MRI slices for different imaging parameters (a) and the independent components (b).

observation that different tissues react differently to the changes of parameters, together with an approximation that each image is composed of linear mixtures of independent tissues, we have used ICA as a mean to isolate each such tissue. This was performed with both simulated and real MRI scans, for various MRI parameter sets. These included the standard 3, to a much enlarged 13 set. The result of applying ICA to the 3 image simulated data can be seen in Fig. 5.4b.

A quick glance to Fig. 5.4b reveals that the independent components show greater tissue selectivity than the original images. The first frame, e.g., depicts nearly only cerebrospinal fluid, whereas the second frame show clearly multiple sclerosis (black spots), invisible in the classification of the raw data.

Tissue segmentation was as well studied on both unprocessed and ICA-processed data, by clustering and judiciously coloring Self-Organizing Maps (SOMs), trained for each data set. The results, for the same simulated data, are shown in Fig. 5.5.



(a) (b) (c) (d) (e)

Figure 5.5: Ground truth segmentation for the simulated MRI data (a), and that found in SOM for the 3-image (b) and the 13-image (c) data sets. Same for ICA-preprocessed in (d) and (e).

Most structures are already segmented in the unprocessed data. Yet, the multiple sclerosis is somewhat hard to detect there, and is only clearly visible from the ICA processed

13-image data set. This suggests that both the increase of multispectral MRI and the pre-processing are efficient strategies for isolating such tissues.

Similar results were observed for real data [1]. Additionally, the application of ICA to the innovation portion of the MR image yields somewhat better results than the ones observed for the normal use of ICA.

ICA preprocessing of MRI data was successfully used in a different segmentation approach, where the acting principle is now a semi-unsupervised use of Support Vector Machines [2]. The results are very promising (see Fig. 5.6).

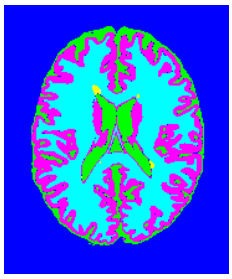


Figure 5.6: Segmentation of the simulated data using support vector machines.

Functional MRI

Functional magnetic resonance imaging (fMRI) is based on similar principles as structural MRI. Yet, with the cost of a clear decrease in spatial resolution, the collection of a full volume in fMRI takes much less time, allowing for multiple images in succession.

Different oxygenation levels in the blood result in variations of its magnetic properties, hence in the MR image. Using this principle, together with a clear pattern of stimulation, several images are collected. One such pattern is the recording of n images with stimulus, followed by a series of m in resting mode. Standard analysis then look for voxels, or combinations of voxels, which show an activation pattern in relation to the stimulus. When ICA is used to process fMRI data, no such stimulus-lock is imposed.

The linear model assumed in our fMRI study is 'transposed' to the one used in EEG and MEG recordings, i.e., the independence is not assumed to exist in time, but rather in space. The mixing matrix, which gives the spatial patterns of activation in EEG and MEG is now the temporal course of the spatially independent components.

Fig. 5.7 shows that, even in a blind source separation framework, ICA is capable of identifying components in clear agreement with the particular auditory type of stimulation. This preliminary study [3] showed as well that brain activity, as well as artifacts, with no relation to the temporal activation of the stimulus can be detected. Due to the lack of relation to the stimulus, those signals would be very hard to identify using classical processing.

Furthermore, even though one asset of ICA is to be unlocked to the stimulus, one may find useful to search for methods that incorporate additional (small) amounts of prior information. These can be based on, *e.g.*, semi-blind source separation 2 or Bayesian methods 3

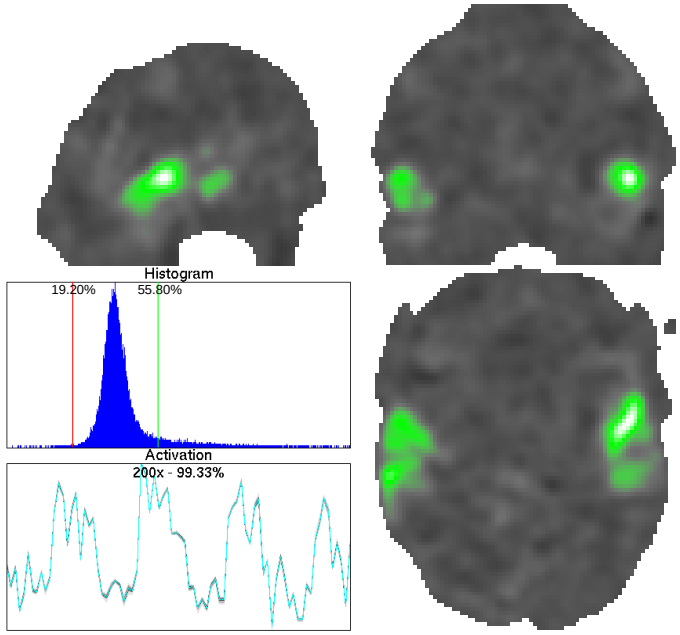


Figure 5.7: Independent component, directly related to the auditory stimulus.

References

[1] Elina Karp, Hugo Gävert, Jaakko Särelä, and Ricardo Vigário. Independent component analysis decomposition of structural MRI. In *Proc. of the 2nd IAESTED Int. Conf. on Biomed. Eng. (BioMED'04)*, Innsbruck, Austria, 2003.

[2] Elina Karp and Ricardo Vigário. Unsupervised MRI tissue classification by support vector machines. In *Proc. of the 2nd IAESTED Int. Conf. on Biomed. Eng. (BioMED'04)*, Innsbruck, Austria, 2003.

[3] Jarkko Ylipaavalniemi and Ricardo Vigário. ICA decomposition of an auditory functional MRI reveals thalamic activation. Submitted to a conference.

Chapter 6

Image analysis applications

Erkki Oja, Jorma Laaksonen, Jukka Iivarinen, Markus Koskela, Ramūnas Girdziušas, Jussi Pakkanen, Ville Viitaniemi, Mika Rummukainen, Mats Sjöberg

6.1 Content-based image retrieval by self-organizing maps

Erkki Oja, Jorma Laaksonen, Markus Koskela, Ville Viitaniemi, Mika Rummukainen, Mats Sjöberg

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 6.1 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.

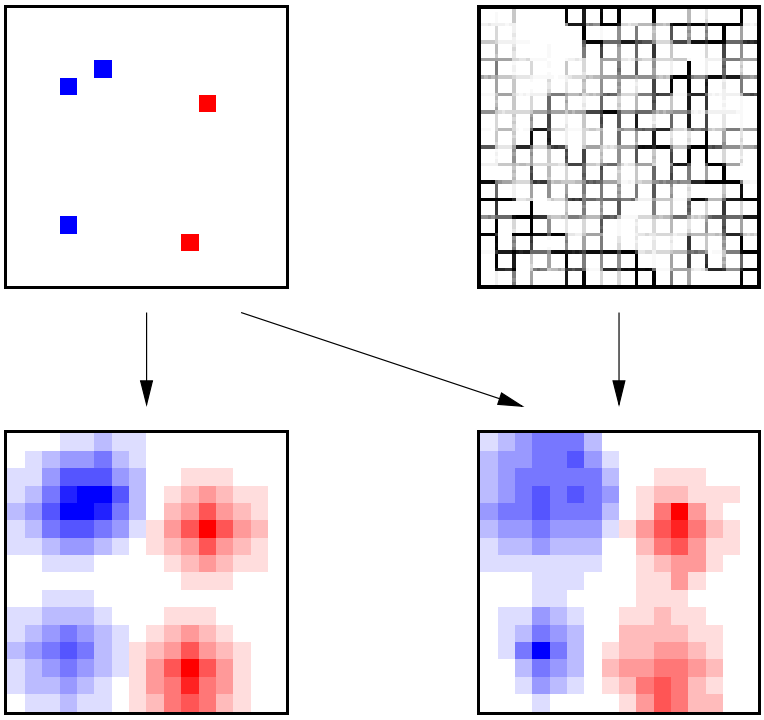


Figure 6.1: An example of how positive and negative map units, shown with red and blue marks on the top-left figure, are low-pass filtered. Two alternative methods exist; either we ignore (bottom-left figure) or take into account (bottom-right figure) the relative distances between neighboring SOM model 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.

MPEG-7 content descriptors

Development of content-based image retrieval techniques has suffered from the lack of standardized ways for describing visual image content. Fortunately, the MPEG-7 international standard has emerged as both a general framework for content description and a collection of specific, agreed-upon content descriptors. MPEG-7 aims at standardizing the description of multimedia content data. It defines a standard set of descriptors that can be used to describe various types of multimedia information. In the scope of our work, the most relevant part of MPEG-7 is the implementation of a set of still image descriptors. Recently, we have integrated the standard MPEG-7 content descriptors into PicSOM [2] and shown that they can be successfully used with it.

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,

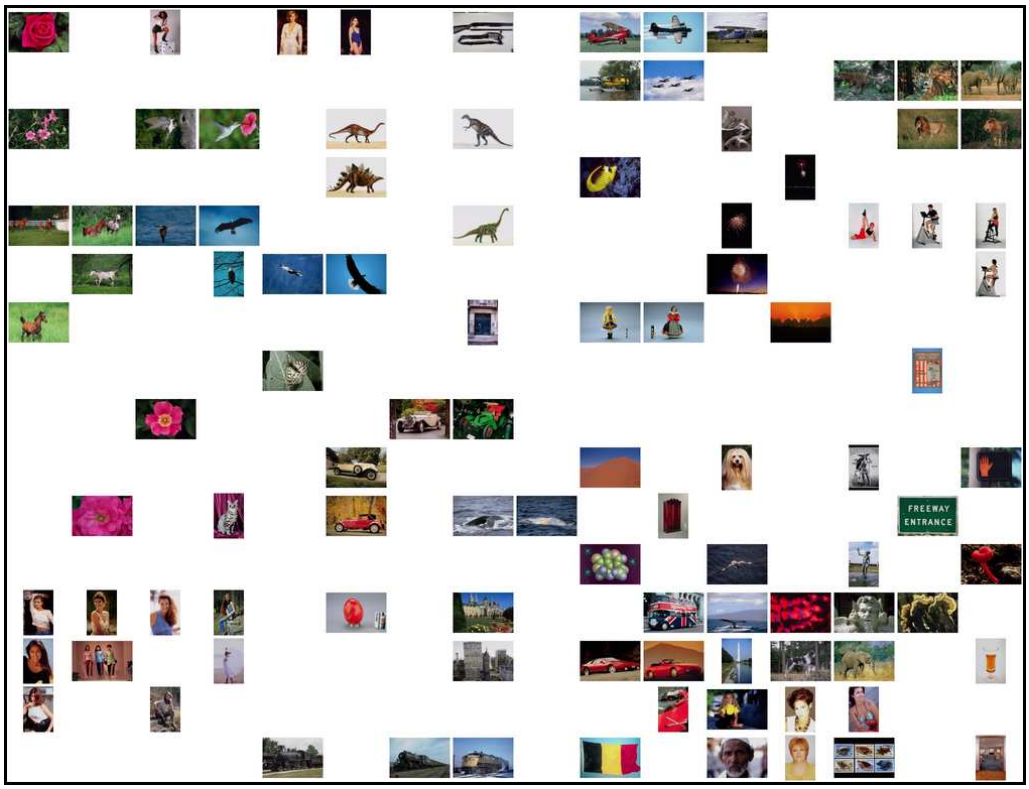


Figure 6.2: The image labels of a 16×16-sized SOM trained with user interaction data.

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 6.2. In the figure, a 16×16-sized SOM trained with user interaction 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, but for an automatic system it 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 implementation of segmentation into PicSOM was done 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.

The results of our preliminary experiments have shown that for most of the used ground truth image classes, the retrieval precision obtained by using both entire and segmented images together excels that obtained by using either ones alone [6]. In a forthcoming series of experiments this will be further ensured.

Application to multimedia messages

We have implemented support for multi-part multimedia messages in the PicSOM system. The system was modified to take advantage of the hierarchical message structure when performing content-based searches. This included implementation of new statistical features for the textual and metadata parts of the messages.

The basic ideas of CBIR can be expanded to more general content-based data retrieval if some kind of low-level statistical feature vectors can be extracted from that particular data. For example, text similarity can be evaluated with the n-gram method. If we also know that some separate data objects are somehow related to each other, we can use this information in the data retrieval. If we, e.g., have a database that contains images of different animals, a short textual description and an audio sample for each of them, we can compare the similarities of the audio samples together with the similarities of the images and texts to obtain the most similar animals as a search result.

We have formed a database of mutually related objects of regular e-mail messages with attachment files, where the message texts and the attachments are probably interrelated. If we now want to search for messages similar to a given reference message, we can use the content-based data retrieval methods of the PicSOM system to first compare similarities of the objects of the same object type in different messages. When these results are then combined, we can evaluate the similarities of the entire messages. On the other hand, instead of searching for whole messages we can use this same approach to help searching for individual message attachments. For example, the text part of a message probably describes what the attachments contain, and the attachments are often related to each other too. 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 other related textual objects. The reference message text probably contains the word “cat”, so when searching for images similar to the one in a reference message, we can also compare the texts of the messages and return images attached to the texts containing similar words as the search result.

CBIR benchmarking

The performance of current CBIR systems is still unclear. The lack of common benchmarks or performance measurement methods and standardized ways of communicating have prevented wide-scale performance comparisons. To overcome these difficulties, researchers have been encouraged to make their CBIR systems compatible with the recently developed Multimedia Retrieval Markup Language (MRML), which aims to be the standard for retrieval system communications. Systems communicating with the same language could then easily attend public contests such as the planned Benchathlon contest (*http://www.benchathlon.net*). Currently, there exists one open source CBIR system, the GNU Image Finding Tool (GIFT) developed at the University of Geneva, that uses MRML.

The MRML language has now been implemented into the PicSOM system which is thus able to communicate with other MRML-based applications. The PicSOM system contains a simple method for benchmarking itself. Now, the benchmarking part also communicates via MRML and this has enabled a performance comparison between PicSOM and GIFT to be run [7]. The results based on the recall-relative precision curves were a little surprising—both CBIR systems managed well in some cases, while at the same time both performed badly in other cases. From the seven classes used in the experiment, the PicSOM and the Separate Normalization algorithm of GIFT ranked first for three image classes, while the CIDF algorithm of GIFT performed best for only one class. These results will be valuable when we will continue the development of the PicSOM system.

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] M. Sjöberg, J. Laaksonen, and V. Viitaniemi. Using Image Segments in PicSOM CBIR System. In *Proc. 13th Scandinavian Conference on Image Analysis*, pages 1106-1113, Halmstad, Sweden, June-July 2003.

[7] M. Rummukainen, J. Laaksonen, and M. Koskela. An efficiency comparison of two content-based image retrieval systems, GIFT and PicSOM. In *Proc. International Conference on Image and Video Retrieval*, pages 500-509, Urbana, IL, July 2003.

6.2 Content-based retrieval of defect images

Jussi Pakkanen, Jukka Iivarinen

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. Currently we use the following features, most of which come from the MPEG-7 standard.

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 2x2-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.

These features were found to work very well on classification experiments using a smaller, pre-classified data base.

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 6.3 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 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.

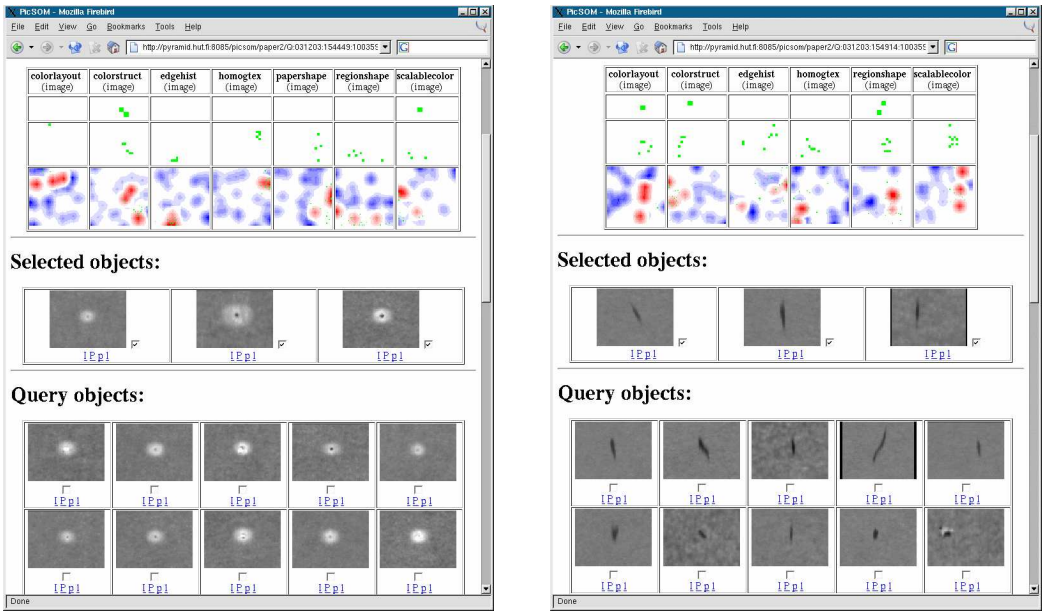


Figure 6.3: Example PicSOM queries.

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.

6.3 Extended fluid-based image registration

Ramūnas Girdziušas, Jorma Laaksonen

Estimation of displacement fields between scalar densities is important in a diverse range of computer vision areas, e.g., patient-to-atlas registration of medical images and object motion field computation.

We investigate a family of state-of-the-art fluid-based image registration (FIR) algorithms posed as a constrained optimization problem [1]. In particular, we analyze the Navier-Stokes prior for the velocity field [2], which depends on Lamé constants μ and λ:

$$\int_{\Omega} \mu ||\nabla \mathbf{v} + \nabla \mathbf{v}^T||^2 + 2\lambda(\nabla \cdot \mathbf{v})^2 \mathbf{dx} \; . \tag{6.1}$$

The choice of the Lamé constants greatly affects image registration results as can be seen in the toy problem shown in Figure 6.4a-e.

There exists evidence that FIR algorithms perform significantly better, provided that the Lamé constants are enriched with spatio-temporal variability. An example is given in Figure 6.4f-g, where at certain points image intensity-driven fluid-registration algorithm produces 5–10 times lower angular errors of the estimated velocity field than one of the state-of-the-art phase-driven optical flow algorithms.

We are constructing an extended FIR model which: (1) behaves stochastically, (2) allows automatic choice of the Lamé constants, (3) utilizes the information from the image registration results at previous time instants.

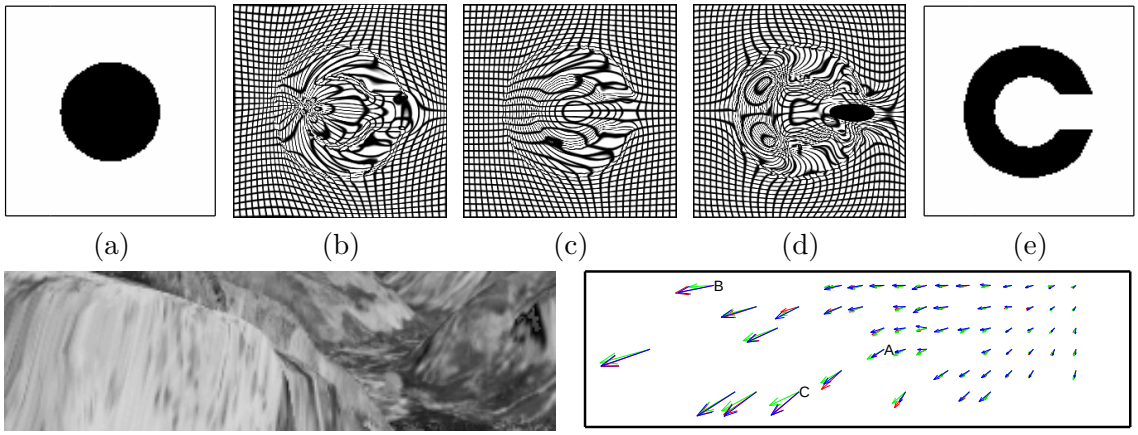


Figure 6.4: Circle (a) matching to letter 'C' (e). Grid deformation is shown: (b) for the weakly elliptic $\lambda + 2\mu \ll 1$ model, (c) Laplacian case $\lambda + \mu = 0$ and (d) strong ellipticity smoothing $\lambda + 2\mu \gg 1$. Optical flow estimation of the simulated 'fly-through' Yosemite valley image sequence. Reference frame (f) and optical flow field (g). True flow is depicted in red color, phase-based estimate is shown in green, and fluid-based results in blue.

References

[1] R. Girdziušas and J. Laaksonen, *Multilayer Perceptron Approach to Non-rigid Image Matching.* ICONIP, Vol. 2, pp.491-496, November, 2001.

[2] G. Christensen, *Deformable shape models for anatomy.* PhD thesis, Washington University, August, 1994.

Chapter 7

On-line recognition of handwritten characters

Vuokko Vuori, Matti Aksela, Jorma Laaksonen, Erkki Oja

7.1 Introduction

Automatic on-line recognition of handwritten text has been an on-going research problem for four decades. It has been gaining more interest lately due to the increasing popularity of hand-held computers, digital notebooks and advanced cellular phones. Traditionally, man-machine communication has been based on keyboard and pointing devices. These methods can be very inconvenient when the machine is only slightly bigger or same size as human palm. Therefore, handwriting recognition is a very attractive input method.

The most prominent problem in handwriting recognition is the vast variation in personal writing styles. There are also differences in one person's writing style depending on the context, mood of the writer and writing situation. The writing style may also evolve with time or practice. A recognition system should be insensitive to minor variations and still be able to distinguish different but sometimes very similar-looking characters. Recognition systems should, at least in the beginning, be able to recognize many writing styles. Such user-independent systems that allow free writing style usually have quite limited recognition accuracies. One way to increase performance is adaptation, which means that the system learns its user's personal writing style.

The goal of the On-line Recognition of Handwritten Characters project has been to develop adaptive methods for on-line recognition of handwritten characters. In this case, adaptation is to be understood in its most demanding sense, i.e. that the system is able to learn new writing styles during its normal use. Due to the learning, the user can use his own natural style of writing instead of some constrained style. Our work has concentrated on recognition of isolated alphanumeric characters and has been carried out in co-operation with Nokia Research Center in years 1997–2002.

The recognition is based on using a set of prototype characters stored in the memory of the system. The input characters are then classified on the basis of their Dynamic Time Warping (DTW) distances to the prototypes. A prototype-based recognition system can easily be adapted to a new writing style by modifying the prototype set: new prototypes can be added, existing prototypes can be reshaped so that they better represent the user's writing style, and prototypes which are not used or which cause more erroneous classifications than correct ones can be inactivated. According to our experiments, best results are obtained if all these three modes of adaptation are used together.

Lately, the character recognizer has been implemented in a Compaq iPAQ PDA device running Linux operating system. Additionally, support for recognition of entire words instead of single characters has been added. This mode is based on a simple language model that uses a dictionary of words. The adaptation of the character prototypes is then carried out after the user has accepted the written word from the given list. Figure 1 depicts this situation: The user has written the characters 'a', 'u', 't', 'o' and the system has recognized each of the characters correctly as shown on the top row of the pop-up list. Also, the corresponding word "auto" has been found in the dictionary along with other words with decreasing similarity.

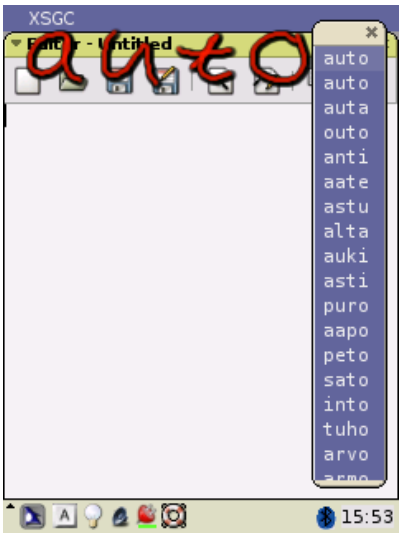


Figure 7.1: The user interface of the character recognition system running in a Linux PDA.

7.2 Adaptive prototype-based character classifiers

Vuokko Vuori, Jorma Laaksonen, Erkki Oja

With adaptive handwritten character recognition systems, it is essential to find a good initial recognition system which performs reasonably well, quickly and accurately, with all kinds of writers. The adaptation process has to be quick in the sense that the user does not have to input several character samples to teach the system a new writing style. In addition, the adaptation should be carried out in a self-supervised fashion during the normal use of the device, i.e. the correct classes of the input characters should be automatically deduced from the user's actions and responses to the recognition results. Naturally, the system should be robust against labelling errors of such training samples.

A prototype set which covers as many as possible alternative ways of writing characters is crucial for the initial recognition system to be able to work well with users using their natural writing styles. We have applied four hierarchical clustering algorithms to a large international database in order to create such a prototype set. In addition, we have experimented with two clustering indices to automatically determine the number of cluster, i.e. different prototypes. On the basis of the results of these experiments, we claim that a good set of prototypes can be formed from the combined results of the different clustering algorithms, but the number of clusters cannot be determined automatically and some human intervention is required [1].

One of the drawbacks of prototype-based classifiers is that the recognition time depends linearly on the size of the prototype set and on the complexity of the similarity measure defined for the prototypes and character samples. The computational complexity of the DTW algorithm depends quadratically on the average number of data points in the prototypes and character samples. We have designed a two-phase recognition scheme in which the prototype set is first pruned and ordered on the basis of a fast preclassification performed with heavily down-sampled character samples and prototypes. Then, the final classification is performed without down-sampling by using the reduced set of prototypes. Faster classification can also be achieved by posing stricter constraints for the nonlinear matching of the data points [1].

Another approach to speed up and enhance the recognition is to prune out those prototypes which are not used by the current user. We performed experiments in which writing styles of several writers were analyzed. The aim of the analysis was to find correlations in the usage of the prototypes and clusters of different writing styles. So the recognition system would be able to predict which prototypes could be pruned on the basis of character samples collected from the user and the estimation of the cluster in which user belongs to. The clustering analysis for the writing styles was performed with a Self-Organizing Map (SOM). The experiments showed that clusters of writing styles can be found, but the writers cannot be reliably assigned to them on the basis of a small set of arbitrary character samples [1].

When prototypes are always added in the adaptation, the recognition rate improves quickly, but the size of the prototype set tends to grow considerably. Therefore the prototype reshaping mode should be utilized too, as the recognition rates will then improve and the size of the prototype set remain the same. However, reshaping is not sufficient when used alone if the user's character samples and the prototypes are too different. According to our experiments, only two new prototypes per class would be enough for adapting the recognition system to a writing style. A prototype inactivation scheme is necessary if some of the samples are incorrectly labeled. Otherwise, the adaptation will be more harmful than useful if the probability of labeling errors is more than approximately 3-4 percent [2].

7.3 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. In addition, 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 [3] 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 [4]. 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 the DEC committee, the classifiers are initialized and ranked in the order of decreasing performance. Results of the member classifiers are used as a one-sided context for the creation of the DEC rules. Each time a character is input to the system, the existing rules are searched through. If no applicable rule is found, the default decision is applied. If the recognition was incorrect, a new rule is created.

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 distributions of distance values from the classifier for each class in every critic. 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 [4].

References

[1] Vuokko Vuori *Adaptive Methods for On-Line Recognition of Isolated Handwritten Characters.* Doctoral Thesis, Helsinki University of Technology, 2002.

[2] Vuokko Vuori, Jorma Laaksonen, and Jari Kangas. Influence of erroneous learning samples on adaptation in on-line handwriting recognition. *Pattern Recognition*, 35(4):915–925, 2002.

[3] Matti Aksela. Comparison of classifier selection methods for improving committee performance. In *Proceedings of MCS2003*, pages 84–93, 2003.

[4] 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.

Chapter 8

Self-organizing map

Teuvo Kohonen, Samuel Kaski, Panu Somervuo, Krista Lagus, Merja Oja, Vesa Paatero

8.1 Self-organizing maps: introduction

Teuvo Kohonen

The name Self-Organizing Map (SOM) signifies a class of neural-network algorithms in the unsupervised-learning category. In its original form the SOM was invented by the founder of the Neural Networks Research Centre, Professor Teuvo Kohonen in 1981-82, and numerous versions, generalizations, accelerated learning schemes, and applications of the SOM have been developed since then.

The central property of the SOM is that it forms a nonlinear projection of a high-dimensional data manifold on a regular, low-dimensional (usually 2D) grid. In the display, the clustering of the data space as well as the metric-topological relations of the data items are clearly visible. If the data items are vectors, the components of which are variables with a definite meaning such as the descriptors of statistical data, or measurements that describe a process, the SOM grid can be used as a groundwork on which each of the variables can be displayed separately using grey-level or pseudocolor coding. This kind of combined display has been found very useful for the understanding of the mutual dependencies between the variables, as well as of the structures of the data set.

The SOM has spread into numerous fields of science and technology as an analysis method. We have compiled a list of over 5000 scientific articles that apply the SOM or otherwise benefit from it.

The most promising fields of application of the SOM seem to be

- data mining at large, in particular visualization of statistical data and document collections,
- process analysis, diagnostics, monitoring, and control,
- biomedical applications, including diagnostic methods and data analysis in bioinformatics, and
- data analysis in commerce, industry, macroeconomics, and finance.

References

[1] Teuvo Kohonen, *Self-Organizing Maps*, Springer Series in Information Sciences, Vol. 30, Springer, Berlin, Heidelberg, New York, 1995, 1997, 2001, 3rd edition.

8.2 5384 works on SOM

Merja Oja, Samuel Kaski, Teuvo Kohonen

The Self-Organizing Map (SOM) algorithm 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. We have collected a comprehensive list of 5384 scientific papers that use the algorithms, have benefited from them, or contain analyses of them. The list is intended to serve as a source for literature surveys.

The collection is available at the WWW address `http://www.cis.hut.fi/nnrc/refs/` (cf. [1, 2]).

A SOM of SOM references. The SOM references were organized onto a document map to study the relationships between the topic categories, and to provide an interface for browsing and searching the collection. A WEBSOM [3] was computed using the titles of the documents. For some documents also an abstract was available and it was used in the computation.

The map is available for browsing and search in the address `http://websom.hut.fi/websom/somref/search.cgi`.

References

[1] Merja Oja, Samuel Kaski, and Teuvo Kohonen. Bibliography of self-organizing map (SOM) papers: 1998–2001 Addendum *Neural Computing Surveys*, Volume 3, pages 1–156, 2003. Available in electronic form at http://www.cse.ucsc.edu/NCS/: Vol 3, pp. 1–156.

[2] Samuel Kaski, Jari Kangas, and Teuvo Kohonen. Bibliography of self-organizing map (SOM) papers: 1981–1997. *Neural Computing Surveys*, 1(3&4):1–176, 1998. Available in electronic form at http://www.cse.ucsc.edu/NCS/: Vol 1, pp. 102–350.

[3] Teuvo Kohonen, Samuel Kaski, Krista Lagus, Jarkko Salojärvi, Jukka Honkela, Vesa Paatero, and Antti Saarela. Self organization of a massive document collection. *IEEE Transactions on Neural Networks*, 11:574–585, 2000.

8.3 Median self-organizing map of human endogenous retroviruses

Merja Oja, Panu Somervuo, Samuel Kaski, Teuvo Kohonen

Only about two percent of human DNA codes for proteins. The function of the rest is unknown, and it has been called "junk DNA." It is, however, far from random, and numerous studies (for a review see [1]) have already shown that it may serve for meaningful functions.

About 45 per cent of the DNA [2] is derived from *transposons*, parts of genome capable of moving or copying themselves in the genome. About eight per cent consists of specific kinds of transposons, 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, 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 [3]. It is important to learn more about the HERVs and their effect on our genome.

We have started studies on human endogenous retroviruses (HERVs) by exploring their mutual relationships and their similarities to other DNA elements [4]. We demonstrated that a completely data-driven grouping is able to reflect same kinds of relationships as more traditional biological classifications and phylogenetic taxonomies. The clusters and their visualization were computed with the Median Self-Organizing Map algorithm [5] of pairwise FASTA-based distances [6]. The whole-sequence distances were able to distinguish between the different known types of endogenous elements, and exogenous retroviruses. The HERVs became grouped meaningfully (see Figure 8.1).

References

[1] Roswitha Löwer. The pathogenic potential of endogenous retroviruses: facts and fantasies. *Trends in Microbiology*, 7(9):350–56, September 1999.

[2] E.S. Lander et al. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

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

[4] Merja Oja, Panu Somervuo, Samuel Kaski, and Teuvo Kohonen. Clustering of human endogenous retrovirus sequences with median self-organizing map. In *WSOM'03 Workshop on Self-Organizing Maps, 9-14 Sep 2003, Hibikino, Japan*, 2003.

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

[6] W. Pearson and D. Lipman. Improved tools for biological sequence comparision. *Proc. Natl. Acad. Sci. USA*, 85:2444–8, 1988.

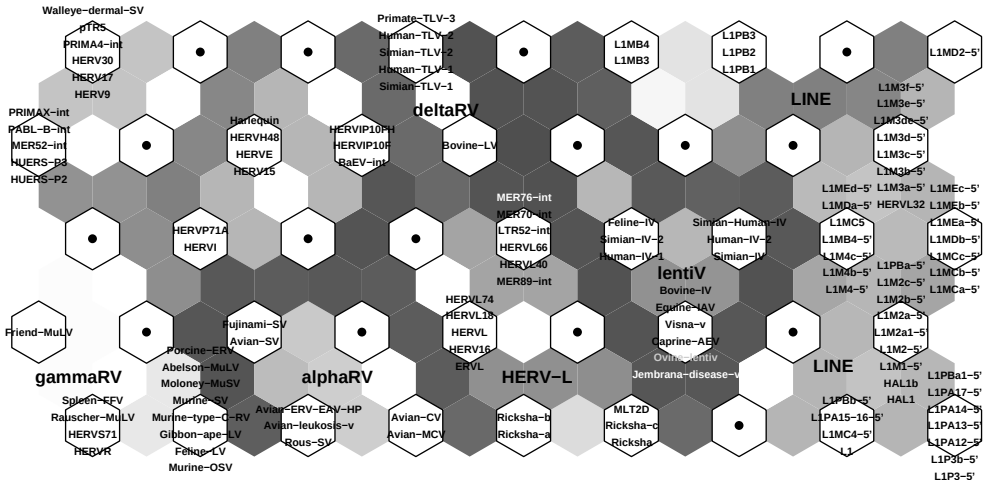


Figure 8.1: Part of the Median SOM of HERV, LINE, and exogenous retrovirus sequences. Every second (bordered, and dotted if not being a best match for any sequence) hexagon denotes a SOM unit, and the rest are U-matrix entries indicating distance between the units. The resulting light areas are clusters and black stripes borders between them. Symbols of the sequences have been inserted to the locations where the sequences have been mapped. Manually assigned names for the clusters are presented on the map. (V=virus, RV=retroV, SV=sarcomaV, OSV=osteoSV, LV=leukemiaV, Mu=murine, TLV=T-lymhocytic V, CV=carcinomaV, AEV=arthitis-encephalitis V, IAV=infectious anemia V, MCV=myelocytomatosis V, FFV=focus forming V.)

8.4 Self-organization of very large document collections

Teuvo Kohonen, Samuel Kaski, Krista Lagus, Vesa Paatero

Text mining systems are developed to aid the users in satisfying their information needs, which may vary from searching answers to well-specified questions to learning more of a scientific discipline. The major tasks of web mining are *searching*, *browsing*, and *visualization*. Searching is best suited for answering specific questions of a well-informed user. Browsing and visualization, on the other hand, are beneficial especially when the information need is more general, or the topic area is new to the user. The SOM, applied to organizing very large document collections, can aid in all the three tasks.

The WEBSOM method

In a method that we have called the WEBSOM [1], a massive collection of documents can be organized efficiently on a large self-organized map.

The computation of document maps. In short, the method is as follows: Encode each document using the vector space model [2] with word weighting. Rare words and a stoplist of common words are excluded. The document vectors are condensed for computational reasons by applying the random projection [3] method. Finally, the document vectors are automatically ordered on a self-organizing map. Various shortcut methods are applied in the construction of large SOMs, including application of a pointer representation in the random projection for fast generation of the document vectors, utilization of the Batch Map algorithm for SOM learning, accelerated winner search, speeded distance computation by neglecting zero-valued elements in the vectors, rapid estimation of a larger map based on a smaller one, and saving memory by using reduced accuracy in storing the maps. It has been shown [1] that while these shortcut methods reduce computation time by an order of $O(d)$ where d is the vocabulary size (nearly 50,000 in our largest experiment), the quality of the maps is practically the same as with a computation where no shortcut methods have been applied.

User interface. The final document map is presented as a series of HTML pages and clickable images that enable exploration of the grid points: a mouse click on a grid point brings to view the links to documents residing in that grid point. The documents, stored in a database, can then be read by following the links. A large map can be first zoomed to view subsets of it more closely. For the largest maps we have used several zooming levels. To provide guidance in the exploration, an automatic method, described in [4], has been utilized for selecting keywords to characterize map regions. The selected words have been marked on the map display.

Content-addressable search. The interface to the map has been provided with a form field into which the user can type a query in the form of a short "document." This query is preprocessed and a document vector is formed in the exactly same manner as for the stored documents. The resulting vector is then compared with the "models" of all SOM grid points, and the best-matching points are marked with circles on the map display: the better the match, the larger the circle. These locations provide good starting points for browsing.

Keyword search. If the user wants to find documents containing a single keyword or very few keywords, one can search the map using a more conventional *keyword search mode* which is provided as an alternative to the content addressable search. The keyword search is performed by accessing an index from each word to the map units where that word occurs.

Experiments

The largest published map

The largest WEBSOM map made so far is a map of 6,840,568 patent abstracts that were available in electronic form and written in English. The size of the SOM was 1,002,240 models (neurons), and the dimensionality of each model was 500. The representation for each document has been made by projecting the 43,222-dimensional word histogram randomly onto the final 500-dimensional space. Formation of the document map and the interface took altogether about 6 weeks with the newest speedup methods; searching occurs in a few seconds.

The Britannica map

For this map, published in [5], the collection consisted of about 68,000 articles from the Encyclopaedia Britannica, and additionally summaries, updates, and other miscellaneous material of about 43,000 items. Very long articles were split into several sections, resulting in a total of about 115,000 documents.

The documents were preprocessed to remove HTML markup, links and images. Inflected word forms were converted to their base forms using a morphological analyzer. The average length of the documents was 490 words. The size of the finally accepted vocabulary was 39,058 words. The words were weighted by the inverse document frequency (IDF). The representation for each document was made by projecting the 39,058-dimensional weighted word histogram randomly onto the final 1000-dimensional space.

The size of the SOM was 12,096 units. Speedups were employed in the creation of the map, namely SOM magnification, and the batch map algorithm where the speeded winner search was employed for fast convergence. The model vectors were represented using reduced accuracy to decrease the memory requirements.

Figure 8.2 exemplifies a case of keyword search. The map and the collection can be explored using a WWW-browser. Further examples of document maps can be found at `http://websom.hut.fi/websom/`.

Once an interesting region has been located e.g. by the search facility, it can be explored by zooming on the map. Figure 8.3 shows an example of how the local ordering of the map may be useful for examining a topic.

Using the document maps for improving search results

Previously, it has been shown how the maps can be utilized for exploration of a large document collection with the help of a browsing interface and the visualized map display. However, as described in [6], the document maps can also be applied for searching without the benefit of the visual interface.

When using the small CISI test collection intended for information retrieval tests, a statistically significant improvement was found when comparing to the standard vector space model. The favourable effect is considered to be due to the fact that the document map brings into the result set similar, relevant documents that do not contain otherwise sufficient amount of the particular words utilized in the search expression.

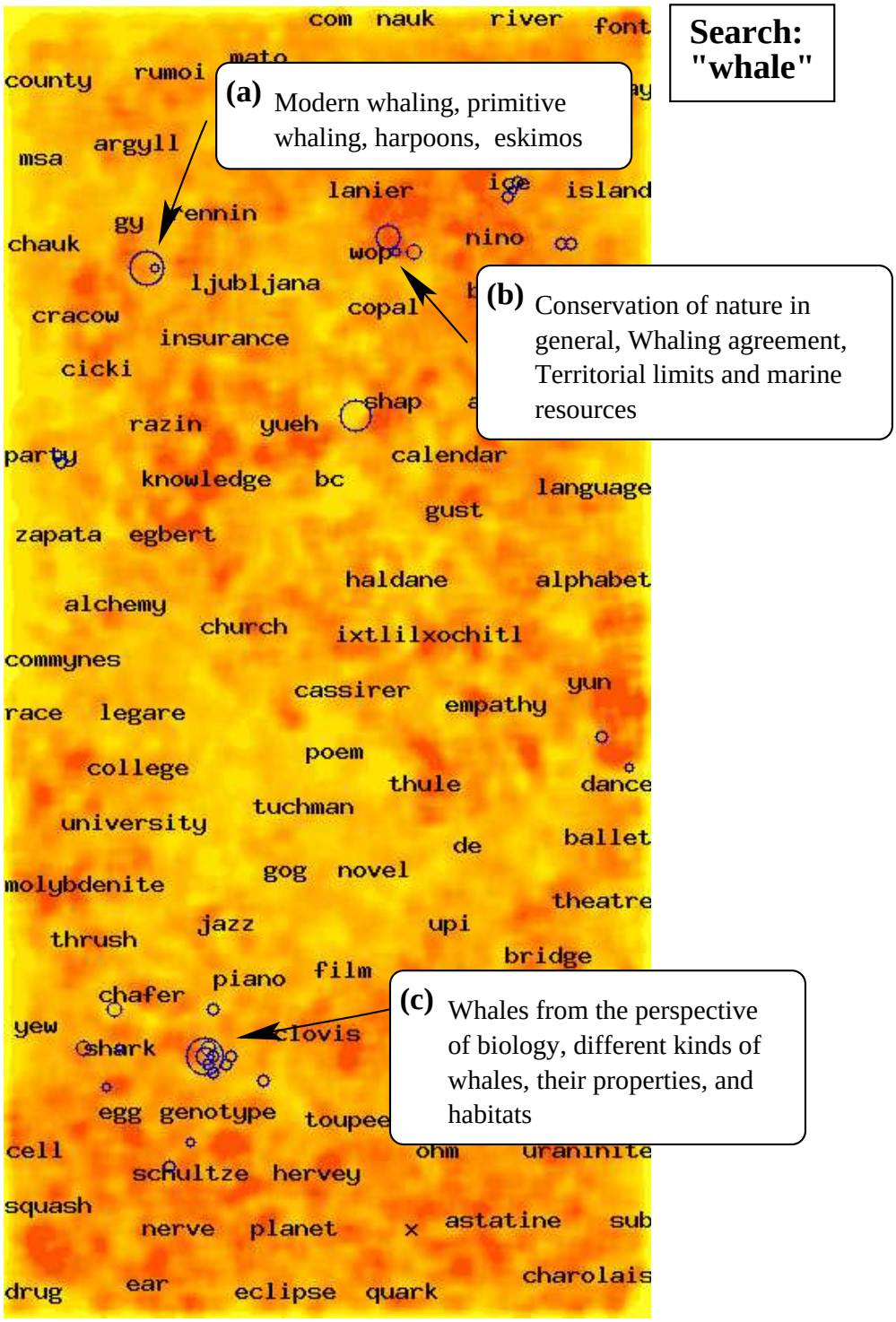


Figure 8.2: The map of Encyclopaedia Britannica articles where the results of a search for 'whale' are depicted. The document map is visualized in the background, and the lighter shades of colour correspond to document clusters. The words written on the document map have been selected automatically using the method described in [4]. The search hits are indicated with blue circles, the size of which describes the goodness of the hit. Three different aspects regarding whales are described in the insets.

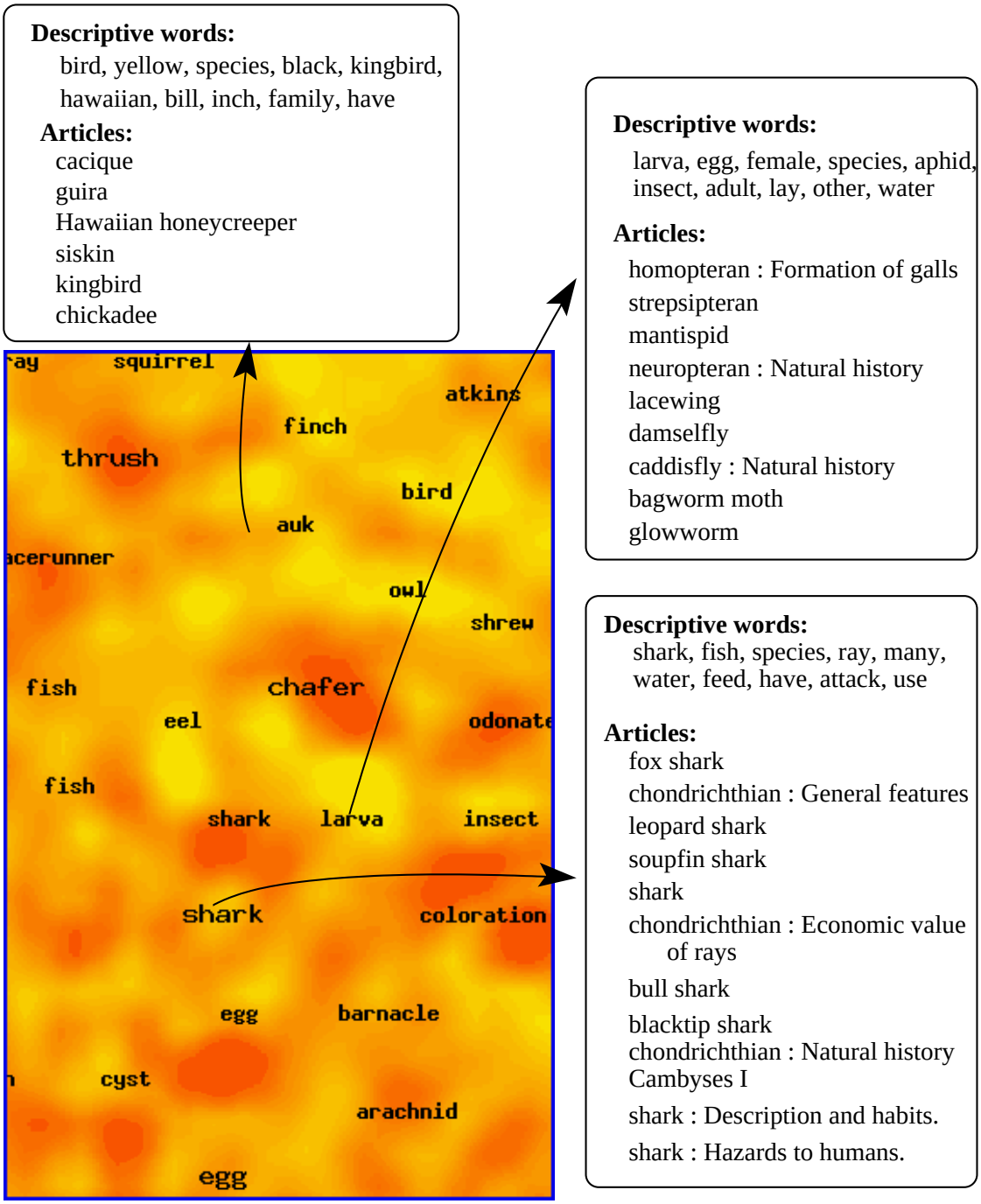


Figure 8.3: A close-up of the map of Encyclopaedia Britannica articles. The user has clicked a map region with the label 'shark', obtaining a view of a section of the map with articles on sharks, various species of fish and eel (in the middle and left); insects and larvae (lower right corner); various species of birds (upper right corner); etc. Searches performed on the map confirm that also whales and dolphins can be found nearby (not shown). A topic of interest is thus displayed in a context of related topics. The three insets depict the contents of three map units, i.e., titles of articles found in the unit. By clicking the title, one may read the article. The 'descriptive words' list was obtained with the labeling method [4] and contains a concise description of the contents of the map unit.

Conclusions

We have demonstrated that it is possible to scale up the SOMs in order to tackle very large-scale problems. The strength of the large map displays is in "finding" rather than "searching for" relevant information. Nevertheless, experiments on a small reference collection indicate that the obtained clusters may serve as meaningful sub-topics that can be used to improve accuracy also in a more focused search task.

Although initially designed for text mining, WEBSOM document maps have additional applications in other fields of natural language processing. Examples of such applications are described in Section 12 where the document maps have been aplied to improving speech recognition and for word sense disambiguation.

References

[1] Kohonen, T., Kaski, S., Lagus, K., Salojärvi, J., Honkela, J., Paatero, V., and Saarela, A. Self Organization of a Massive Document Collection. *IEEE Transactions on Neural Networks*, vol. 11, number 3, pp. 574–585. May 2000.

[2] Salton, G., and McGill, MJ. *Introduction to modern information retrieval*. McGraw-Hill, New York, 1983

[3] Kaski, S. Dimensionality reduction by random mapping. *In Proc of IJCNN'98, Int Joint Conf on Neural Networks.* IEEE Press, Piscataway, NJ, 1998, pp. 413–418

[4] Lagus, K., and Kaski, S. Keyword selection method for characterizing text document maps. In *Proc. of ICANN99, Ninth Int. Conf. on Artificial Neural Networks* vol. 1, pp. 371–376. IEE, London, 1999.

[5] Lagus, K., Kaski, S., and Kohonen, T. Mining massive document collections by the WEBSOM method *Information Sciences*. In press.

[6] Lagus, K. Text retrieval using self-organized document maps. *Neural Processing Letters*, vol. 15, no. 1, pp. 21-29. February 2002.

Chapter 9

Adaptive cognitive systems

Timo Honkela, Aapo Hyvärinen, Krista Lagus, Ville Könönen, Kevin I. Hynnä, Juha Winter, Jaakko Väyrynen

9.1 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, and constructive learning. We also work in close collaboration with other groups in our laboratory, e.g., related to multimodal environments and sensory fusion.

9.2 Unsupervised learning for agent communication

Traditional cognitive models and language technologies widely neglect the fact that language users learn the language in a large varity of contexts. This leads into varying interpretation of expressions. When this aspect is not carefully considered also the practical applications suffer from problems related to the basic underlying assumptions. Perhaps the most striking classical example can be given in the area of information retrieval. Namely, it has been found that in spontaneous word choice for objects in five domains, two people favored the same term with less than 20% probability. Moreover, it has been shown that different indexers, well trained in an indexing scheme, might assign index terms for a given document differently. It has also been observed that an indexer might use different terms for the same document at different times. In summary, while developing models of linguistic cognition or some computational tools, we cannot assume that there is a commonly shared model among the language users. On contrary, we have to be able to develop systems that are capable, among other things, to conduct meaning negotiations.

Abstract model of adaptive communicating agents

At an advanced level of multi-agent co-operation, as mentioned above, each agent has its own model of the environment. Thus, each agent has an individual interpretation for the relationship between the messages and the environment. These differences in the agents' models motivate the development of methods which provide the agents with the ability to learn, including learning to interpret messages from other agents.

The agents can perceive their environment, they are part of it, and possibly they can change it. The environment may be a computerized representation, constructed, or natural. The borderlines of these domains may, of course, be vague. A natural environment, in particular, is ever-changing, and consists of various continuous phenomena.

We have considered the possibility of applying a natural, or near-natural language as the communication medium. The general properties of natural languages necessitate some capabilities that autonomous agents will need to have. These basic properties of natural languages and their interpretation include ambiguity, contextuality, open-endedness, vagueness, and subjectivity.

Ambiguity or vagueness, then, can be considered a necessity when the communication medium is used in an open and changing environment in which having a distinct and a priori determined symbol, or combination of symbols, would be difficult, or even impossible. Finally, to ensure successful communication, both the sending and the receiving agent must share a similar enough framework of interpretation, and the message or the situation must contain enough information to activate a proper framework for the receiver.

Adaptive communicating agents based on Self-Organizing Map

We have developed a model of communicating agents based on the self-organizing map algorithm [5]. An agent has the following properties: it can perceive its simulated environment, it can move in its environment, it can perform some simple actions, and it can send and receive messages. The main components of its internal structure include a short-term and a long-term memory, and a decision making mechanism. Although these two memory types are closely interconnected, they have different implementations: episodic memory is dynamic and accurate in nature, whereas semantic memory is adaptive and approximative being based on the self-organizing map. The key idea is to provide the means for each agent to associate continuous-valued parameter spaces to sets of symbols, and furthermore, to "be aware" of the differences in this association and to learn those differences explicitly.

These kinds of abilities are especially required by highly autonomous agents that need to communicate using an open set of symbols or constructs in the agent language. [2,4]

The self-organizing map is especially suitable for the central processing element of autonomous agents because of the following reasons:

- The self-organizing map algorithm modifies its internal presentation, i.e., the codebook vectors, according to the external input which enables the adaptation of the agents.
- The self-organizing map is able to process natural language input to form "semantic maps" [6].
- Symbols and continuous variables may be combined in the input, and are associated by the self-organizing map. Continuous variables may be quantized, and a symbolic interpretation can be given for each section in the possibly very high-dimensional space of perceptual variables [1].
- Because the self-organizing map implements unsupervised learning, processing external input without any prior classifications is possible. The autonomous agent may form an individual model of the environment and of the relation between the expressions of the language and the environment [2]. In general, the basic approach is compatible with the framework of constructive learning theories [3].

References

[1] T. Honkela. Self-Organizing Maps in Symbol Processing. In *Hybrid Neural Systems*, Stefan Wermter, Ron Sun (eds.), Springer, Heidelberg, 2000, pp. 348-362.

[2] T. Honkela, K.I. Hynnä, and T. Knuuttila. Framework for Modeling Partial Conceptual Autonomy of Adaptive and Communicating Agents. *Proceedings of Cognitive Science*, 2003.

[3] T. Honkela, T. Leinonen, K. Lonka, and A. Raike. Self-Organizing Maps and Constructive Learning. *Proceedings of ICEUT'2000, International Conference on Educational Uses of Communication and Information Technologies*, Beijing, China, August 21-25, 2000, pp. 339-343.

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

[5] T. Kohonen. *Self-Organizing Maps*. Springer, 2001.

[6] H. Ritter, and T. Kohonen. Self-Organizing Semantic Maps. *Biological Cybernetics*, 61:241-254, 1989.

9.3 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 Markovian 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 Markovian property does not hold and more advanced methods should be used instead. A powerful tool for handling these highly non-Markovian 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*. In a Markov game, the process changes its state according to the action choices of all agents and can thus be seen as a multicontroller MDP. In Fig. 9.1, there is an example of a Markov game 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.

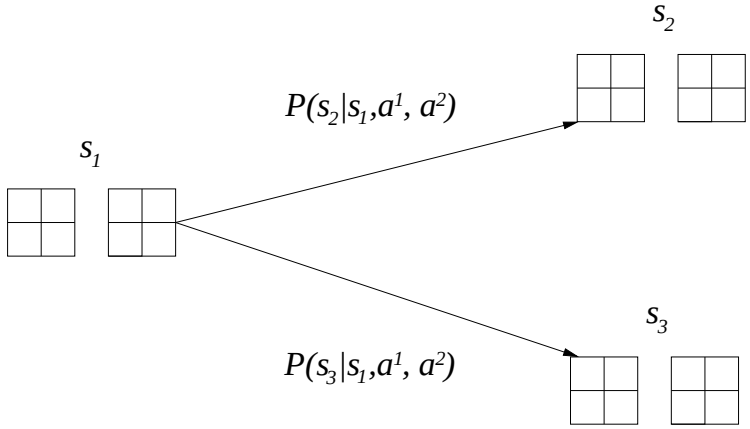


Figure 9.1: An example Markov game with three states.

In single-agent MDPs, it suffices to maximize the utility of the agent in each state. In Markov games, 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:

$$Q^i_{t+1}(s_t, a^1_t, \ldots, a^N_t) = (1 - \alpha_t)Q^i_t(s_t, a^1_t, \ldots, a^N_t) + \alpha_t[r^i_{t+1} + \gamma f(s_{t+1})], \qquad (9.1)$$

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 $a^1_t, \ldots, a^N_t$. 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].

Grid world example

In this section we provide a simple example of multiagent reinforcement learning. Let us consider a grid world containing nine cells, two competing agents and two goal cells (Fig. 9.2). Initial positions of the agents are the bottom corners 1 and 2, respectively, and they can move to adjacent cells (4-neighborhood) on each round. An agent gets a large positive payoff when it founds the right goal cell. Additionally, it gets a small negative payoff if it collides with its opponent, i.e. both agents move into the same cell, and the agents are returned back to their original cells.

G2		**G1**
1		**2**

Figure 9.2: The game board used in the grid world example. Agents are initially located in the cells marked with numbers **1** and **2**. Goal cells are marked with symbols **G1** and **G2**.

When this problem is modeled as a Markov game, a state is a pair containing the positions of the agents and the actions are the directions of movement. The problem was solved by using the asymmetric multiagent reinforcement learning method with discount factor $\gamma = 0.99$. In this asymmetric setting, the agent 1 decides his action first (leader) and the agent 2 (follower) reacts optimally to this selection. The learning process converged to the optimal paths (policy functions) shown in Fig. 9.3. Corresponding convergence curves can be found in Fig. 9.4, in which changes in Q-values, i.e. the Euclidean distance between two vectors containing Q-values of the consecutive iterations of the learning algorithm, are plotted against iteration rounds. More detailed empirical evaluations of the asymmetric learning method can be found in [4] and [5].

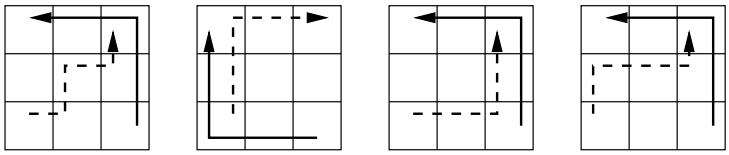


Figure 9.3: Some optimal paths generated by the asymmetric multiagent reinforcement learning model.

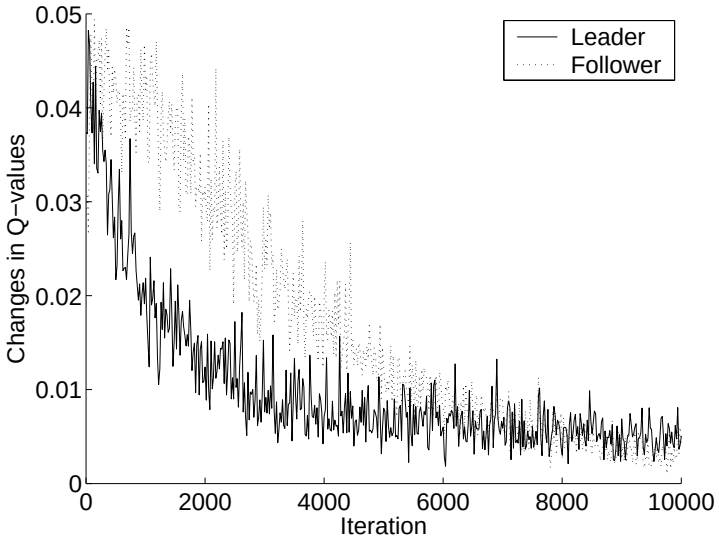


Figure 9.4: The convergence of the asymmetric learning method in the grid world example problem.

References

[1] A. Greenwald and K. Hall. Correlated-Q learning. In *Proceedings of the AAAI-2002 Spring Symposium Workshop on Collaborative Learning Agents*, Stanford, CA, 2002. AAAI Press.

[2] J. Hu and M. Wellman. Multiagent reinforcement learning: Theoretical framework and an algorithm. In *Proceedings of the Fifteenth International Conference on Machine Learning (ICML'98)*, Madison, WI, 1998. Morgan Kaufmann Publishers.

[3] V. Könönen. Asymmetric multiagent reinforcement learning. In *Proceedings of the 2003 WIC International Conference on Intelligent Agent Technology (IAT-2003)*, Halifax, Canada, 2003. IEEE Press.

[4] V. Könönen. Gradient based method for symmetric and asymmetric multiagent reinforcement learning. In *Proceedings of the Fourth International Conference on Intelligent Data Engineering and Automated Learning (IDEAL 2003)*, Hong Kong, China, 2003. Springer-Verlag.

[5] V. Könönen. Policy gradient method for multiagent reinforcement learning. In *Proceedings of the 2nd International Conference on Computational Intelligence, Robotics and Autonomous Systems (CIRAS 2003)*, Singapore, 2003.

9.4 Emergence of linguistic features using Independent Component Analysis

Language technology is very central area in the development of intelligent systems. Traditionally, much of the development work has been manual: encoding linguistic and domain knowledge even for a single system may even take years. We have been studying how aspects of human language learning could be modeled, potentially resulting into (a) more realistic cognitive models than those based on, e.g., rule-based representations, and (b) efficient tools for applications in, for instance, information retrieval, natural language interfaces, machine translation and computer supported collaborative work. In the following, we consider one specific aspect of language learning, i.e, how categories of words can be learned from input data without supervision.

Word Category Learning

A word can belong to several syntactic categories simultaneously. The number of categories is even higher if one takes into account the semantic categories. Such categorization has traditionally been determined by hand: the categories into which a word belongs to are described in a manually collected dictionary.

In order to facilitate learning of word categories, the self-organizing map has earlier been used in the analysis of word context data, e.g., in [4] (artificially generated short sentences), and [1] (Grimm fairy tales). The result of a word context analysis based on the self-organizing map algorithm can be called a word category map. Areas or local regions on a word category map are implicit categories that have emerged during the learning process. Single nodes in the map can be considered as adaptive prototypes. Each prototype is involved in the adaptation process in which the neighbors influence each other and the map is gradually finding a form in which it can best represent the input.

One classical approach for defining concepts is based on the idea that a concept can be characterized by a set of defining attributes. In contrast, the prototype theory of concepts involves that concepts have a prototype structure and there is no delimiting set of necessary and sufficient conditions for determining category membership that can also be fuzzy. Instances of a concept can be ranked in terms of their typicality. Membership in a category is determined by the similarity of an object's attributes to the category's prototype.

The emergent categories on a word category map are implicit. The borderlines for any categories have to be determined separately. It would be beneficial if one could find more explicit categories in an automated analysis. Moreover, each word appears in one location of the map. This means, among other things, that one cannot have a map in which several characteristics or categories of one word would be represented unless the the categories overlap and accordingly the corresponding areas of the map overlap. In some cases, this is the case: it is possible to see the area of modal verbs inside the area of verbs, e.g., on the map in our earlier research [1]. However, one might wish to find a sparse encoding of the words in such a way that there would be a collection of features associated with each word. For instance, a word can be a verb, a copula and in past tense. It is an old idea in linguistics to associate words with features. The features can be syntactic as well as semantic However, these features are, as already mentioned, given by hand, and the membership is crisp.

Independent component analysis of word contexts

We have studied the emergence of linguistic representations through the analysis of words in contexts using the Independent Component Analysis (ICA) [3]. 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. As a simple example, the method is able to find a feature that is shared by words such as "must", "can" and "may", i.e. modal verbs.

In one of our experiments, we formed a context matrix $\mathbf{C}$ in which c_{ij} denotes the number of occurrences jth word in the immediate context of ith word, i.e, ith word followed by jth word with no words between them. This provided a 100×2000 matrix. A logarithm of the number of occurrences was taken in order a reduce the effect of the very most common words in the analysis and finally each word vector was normalized to unit length.

The results of the ICA analysis corresponded in most cases very well or at least reasonably well with our preliminary intuitions. The system was able to automatically create distributed representations as a meaningful collection of emergent linguistic features; each independent component was one such feature. For instance, Fig. 9.5 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. 9.6. The reason appears to be that "psychology", "neuroscience", and "science" share a semantic feature of being a science or a scientific discipline. This group of words provide a clear example of distributed representation where, in this case, two components are involved.

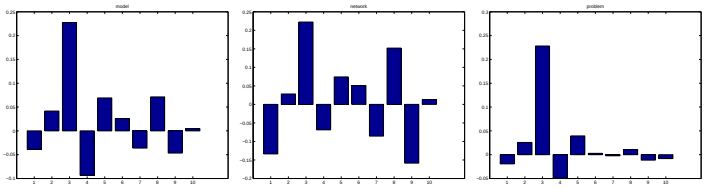


Figure 9.5: ICA features for "model", "network" and "problem".

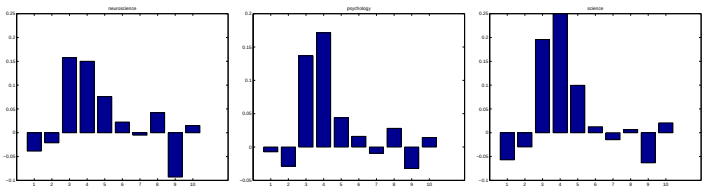


Figure 9.6: ICA features for "neuroscience", "psychology" and "science".

We have been able to show how independent component analysis can bring an additional advantage of finding explicit features that characterize words in an intuitively appealing manner. Independent component analysis appears to make possible a qualitatively new kind of result which have earlier been obtainable only through hand-made analysis. In the future, we will study more in detail what is the relationship between automatically acquired categories and manually defined ones.

References

[1] T. Honkela, V. Pulkki, and T. Kohonen. Contextual Relations of Words in Grimm Tales, Analyzed by Self-Organizing Map. Proceedings of ICANN'95, International Conference on Artificial Neural Networks, Paris, France, October 9-13, 1995, vol. 2, pp. 3-7.

[2] T. Honkela, A. Hyvärinen, and J. Väyrynen. Emergence of Linguistic Features using Independent Component Analysis. Helsinki University of Technology, Publications in Computer and Information Science Report, 2003.

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

[4] H. Ritter, and T. Kohonen. Self-Organizing Semantic Maps. *Biological Cybernetics*, 61:241-254, 1989.

Chapter 10

Bioinformatics

Samuel Kaski, Janne Nikkilä, Merja Oja, Leo Lahti, Jarkko Venna, Eerika Savia, Janne Sinkkonen, Jaakko Peltonen

10.1 Introduction

Bioinformatics refers to the study of biomedical data using methods from mathematics, statistics, and computer science. In particular, gene sequencing and functional genomics experiments produce massive amounts of high-dimensional data that are being collected into community resource databases. 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.

Mining the data to generate hypotheses about gene function and regulation is the next big challenge. Statistical machine learning and mining methods can contribute by learning flexible models of regularities in data. Our research has had two interlinked goals: (1) to develop and apply information visualization and clustering methods for exploring the functional genomics data sets, and (2) to develop methods for focusing the analysis to interesting variation in data. The key assumption is that structures appearing in several data sets are more relevant, and hence dependencies between data sets may reveal interesting findings. Methods of learning metrics and dependency modeling (Section 11) will be used.

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.

First steps of a new project in analyzing human endogenous retroviruses were described in Section 8.3.

10.2 Exploratory analysis of gene expression

Exploratory analysis is an irreplaceable first step in bioinformatics research, in particular of gene expression data. Interesting findings need to be made amidst the unknown interactions of thousands of genes, and distinguished from biological and measurement noise.

Visualization plays an important role in the exploratory analysis. The Self-Organizing Map (SOM), developed in the Neural Networks Research Centre, is particularly useful since it constructs a nonlinear projection of the data to a map display which can be used for visualizing of similarity relationships and cluster structures. The SOM has been used successfully to generate hypotheses about regularities in gene expression data [1, 2, 3]. Figure 10.1 shows an example where a SOM-based display revealed the density structure in yeast *Saccharomyces cerevisiae* gene expression measurements [4]. As a result, the biological experiments producing locally the most variation could be discerned. Moreover, we found functionally meaningful subsets of genes, for example ribosomal proteins, and also confirmed that clsses of an existing functional classification are homogeneous in terms of their expression.

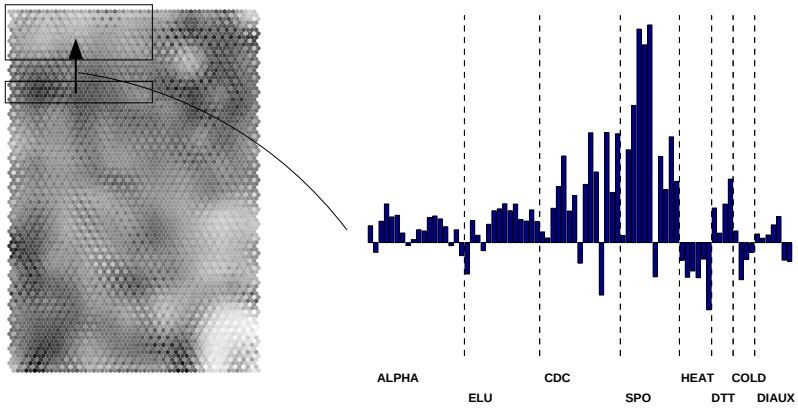


Figure 10.1: Left: SOM representing yeast gene expressions. Right: Difference between data in the cluster in the upper left corner and the area below it. Strong expression in the sporulation (SPO) is the main characteristic distinguishing the cluster.

A key question in visualization is the preservation of the original similarity relationships. In general, it is impossible to preserve all the similarites 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. We argue that, for data exploration, it is more important that the visualizations are trustworthy.

We studied the trustworthiness of SOM and other visualization methods with gene expression data [1, 3]. The SOM was found to be more trustworthy than other methods, except for the smallest neighborhoods (Figure 10.2).

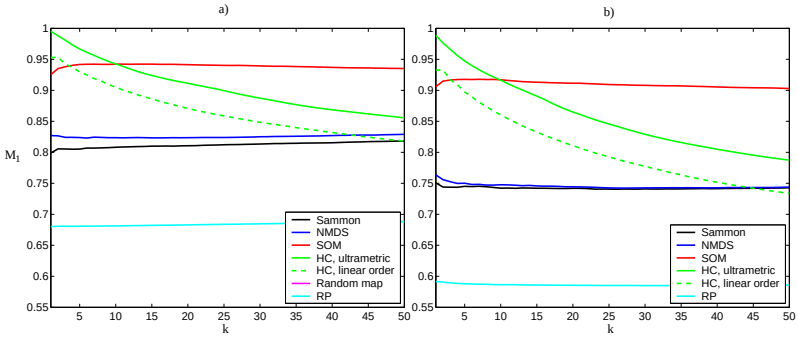


Figure 10.2: Trustworthiness of the visualized similarities (neighborhoods of k nearest samples) for methods that visualize similarity relationships in data. Sammon: Sammon's mapping, NMDS: non-metric multidimensional scaling, SOM: self-organizing map, HC: hierarchical clustering, with the ultrametric distance measure and with the linear distance measure. RP: Random linear projection is the approximate worst possible practical result (the small standard deviation over different projections, approximately 0.01, is not shown). The theoretical worst case, estimated with random neighborhoods, is approximately $M_1 = 0.5$. **a)** Yeast data. **b)** Mouse data.

10.3 Exploratory analysis of dependencies between functional genomics data sets

Exploratory analysis can be enhanced by focusing on relevant variation in the data set, the relevance being determined by another, auxiliary data set. Dependency modeling and learning metrics methods (Section 11) provide a state of the art tool for this, and we have developed and applied them for visualization and clustering in bioinformatics problems.

Visualizations by Maximizing Dependency

A main problem in gene expression analysis is the correct choice of similarity measure, or the metric. It can be learned automatically with the *learning metrics principle* (see Section 11).

We have visualized similarity relationships of gene expression profiles with SOMs in learning metrics. For instance, the metric used in visualizing yeast gene expression was supervised by functional classes of the genes. Visualization of human gene expression was supervised by better-known homologous mouse genes.

Alternatively, expression data can be visualized with a linear projection that generalizes classical methods (Section 11). The results of supervising the projection by functional classes of yeast genes suggest that most functional classes are not strongly differentiated in the expression data, while some information about the overlap of the classes and about their division into subclasses can be found (Fig. 10.3).

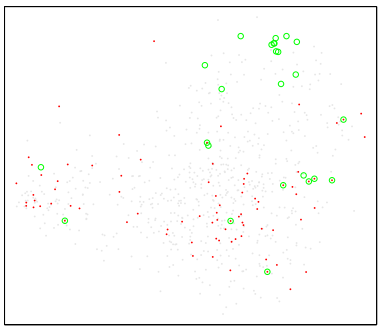


Figure 10.3: Yeast gene expression profiles projected onto two informative components, with protein synthesis (green circle) and mitochondrial organization (red dot) functional classes shown. The protein synthesis class has a subclass at the top.

Clusters by Maximizing Dependency

Dependency maximizing clustering methods (Section 11) are a principled way of finding dependencies between data sets, and presenting them in the form of clusters.

Discriminative clustering (DC) (Section 11) was applied to search for yeast stress genes [7]. We tested the method by replicating the findings of an earlier study. Stress genes should be active in all stress treatments, and additionally potentially 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 should focus the clusters on behavior regulated by MSN2/4.

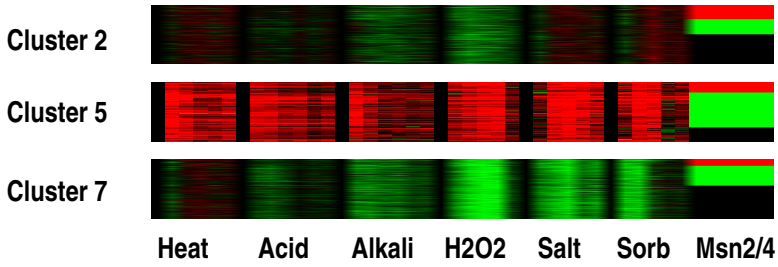


Figure 10.4: Discriminative clusters revealing a group of yeast stress genes (cluster 5) that are putatively regulated by transcription factors MSN2 and MSN4, that is, they are upregulated normally, but downregulated when *Msn2* and *Msn4* are mutated to non-functional. The six leftmost columns are the gene expressions of unmutated yeast and the rightmost column is the discretized gene expression of mutated yeast under stress. Clusters 2 and 7 are examples of an expression cluster without dependency to transcription factors. (Red = upregulation, green= downregulation).

We identified a subset of genes that are upregulated in all stress conditions, but only when regulators MSN2 and MSN4 are functional. Figure 10.4 presents both the gene expressions of normal yeast and the discretized expression of the mutated yeast genes in all DC clusters. Stress genes found in an independent study were strongly enriched in the discovered subset.

Yeast gene regulatory mechanisms were explored with *associative clustering (AC)* (Section 11), by searching for gene groups that are maximally dependent by expression [5] and by transcription factor binding [8]. We found statistically significant dependency, confirmed the results with known regulatory mechanisms, and generated hypotheses for new regulatory interactions.

In a novel application we explored the dependency between expression of human and mouse genes that are putatively orthologous, that is, similar by their sequences [6]. Associative clustering summarizes the data as sets with regularities in their behavior in the two organisms, and outlier sets (Fig. 10.5).

References

[1] Janne Nikkilä, Petri Törönen, Samuel Kaski, Jarkko Venna, Eero Castrén, and Garry Wong. Analysis and visualization of gene expression data using self-organizing maps. *Neural Networks*, 15:953–966, 2002.

[2] Merja Oja, Janne Nikkilä, Petri Törönen, Garry Wong, Eero Castrén, and Samuel Kaski. Exploratory clustering of gene expression profiles of mutated yeast strains. In Wei Zhang and Ilya Shmulevich, editors, *Computational and Statistical Approaches to Genomics*, pages 65–78. Kluwer, Boston, MA, 2002.

[3] 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.

[4] Michael B. Eisen, Paul T. Spellman, Patrick O. Brown, and David Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95:14863–14868, 1998.

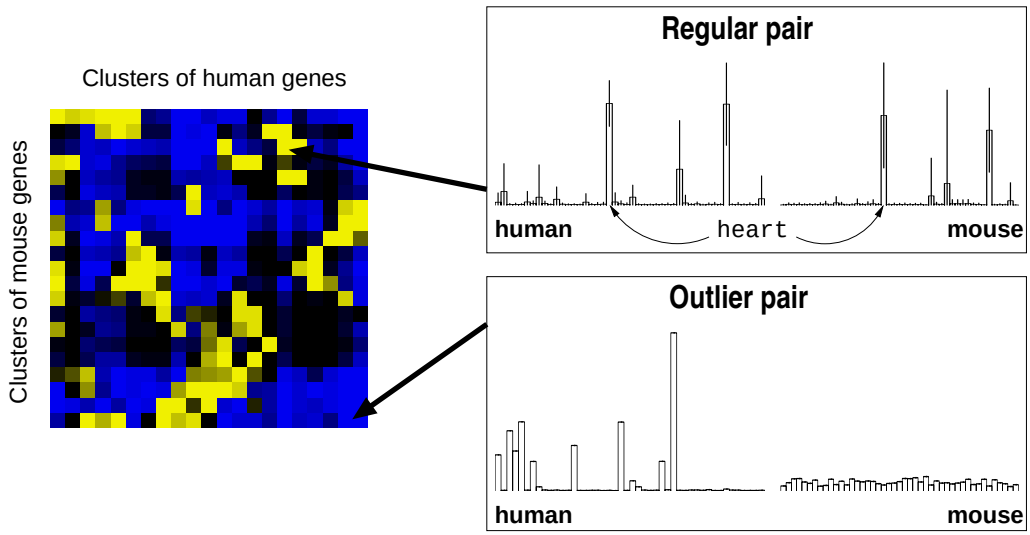


Figure 10.5: Contingency table from AC trained with gene expressions of orthologous genes of human and mouse, that reveals unexpectedly common (yellow) gene pairs and unexpectedly rare gene pairs (blue). An example of both cases is given. **Regular pair:** orthologous genes are expressed strongly in the same tissue, heart, in both organisms. **Outlier pair:** A possibly interesting case where the gene is expressed in human but not at all in mouse, due to functional differences or measurement errors.

[5] Timothy R. Hughes, Matthew J. Marton, Allan R. Jones, Christopher J. Roberts, Roland Stoughton, Christopher D. Armour, Holly A. Bennett, Ernest Coffrey, Hongyue Dai, Yudong D. He, Matthew J. Kidd, Amy M. King, Michael R. Meyer, David Slade, Pek Y. Lum, Sergey B. Stepaniants, Daniel D. Shoemaker, Daniel Gachotte, Kalpana Chakraburtty, Julian Simon, Martin Bard, and Stephen H. Friend. Functional discovery via a compendium of expression profiles. *Cell*, 102:109–126, 2000.

[6] Andrew I. Su, Michael P. Cooke, Keith A. Ching, Yaron Hakak, John R. Walker, Tim Wiltshire, Anthony P. Orth, Raquel G. Vega, Lisa M. Sapinoso, Aziz Moqrich, Ardem Patapoutian, Garret M. Hampton, Peter G. Schultz, and John B. Hogenesch. Large-scale analysis of the human and mouse transcriptomes. *PNAS*, 99:4465–4470, 2002.

[7] Helen C. Causton, Bing Ren, Sang Seok Koh, Christopher T. Harbison, Alanita Kanin, Ezra G. Jennings, Tong Ihn Lee, Heather L. True, Eric S. Lander, and Richard A. Young. Remodeling of yeast genome expression in response to environmental changes. *Molecular Biology of Cell*, 12:323–337, February 2001.

[8] 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.

Chapter 11

Learning metrics

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

11.1 Introduction

Unsupervised learning such as clustering and information visualization suffers from the garbage in—garbage out problem. The ultimate goal is to make discoveries in data, that is, to find new things without speficying them in advance. The problem is that unsupervised learning cannot distinguish relevant variation from irrelevant variation in data. Structured noise becomes modeled as well as relevant structure.

Hence, all successful unsupervised learning must have been supervised implicitly or explicitly, by feature extraction or model selection. Our goal is to automate (part of) this implicit supervision by learning from a supervising signal. The difference from standard supervised learning is that the goal is to explore new things in the primary data given the supervision, whereas in supervised learning the goal is simply to predict the supervisory signal. The task could be coined supervised unsupervised learning.

Sample applications include exploration of factors leading to bankruptcy, where primary data are financial indicators and supervisory signal is the bankruptcy risk. Another is exploration of gene expression, supervised by functional classes of the genes.

For methods that are based on distance computations, the supervision can be conveniently incorporated in the distance measure. The idea of deriving information-geometric metrics to data spaces from paired data has been coined the learning metrics principle. It is assumed that variation of the primary data $\mathbf{x} \in \mathbb{R}^n$ is important only to the extent it causes variation in *auxiliary data* c, the supervisory signal, which is available paired to the primary data.

In other words, important variation in $\mathbf{x}$ is supposed to be revealed locally by variation in the conditional density $p(c|\mathbf{x})$. 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.

$$d_L^2(\mathbf{x}, \mathbf{x} + d\mathbf{x}) \equiv D_{\mathrm{KL}}(p(c|\mathbf{x}) \| p(c|\mathbf{x} + d\mathbf{x})) = \; d\mathbf{x}^T \mathbf{J}(\mathbf{x}) d\mathbf{x} \; , \tag{11.1}$$

where $\mathbf{J}(\mathbf{x})$ is the Fisher information matrix. The Riemannian metric depends on $\mathbf{x}$ and hence is more general than a global scaling of the feature space.

The Fisher information matrix has earlier been used to construct metrics to spaces of probability models (see, e.g., [1]). The novelty here is that the information matrix is applied in the data space to construct a new metric there. The coordinates of data are considered as parameters

In practice, the idea 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 these approaches are discussed in more detail below.

11.2 Learning metrics for information visualization

Explicit estimation of learning metrics by approximations to (11.1) is generally applicable to explicitly supervise unsupervised metric-based methods. The choice of auxiliary data determines what is important, without need for hand-tuned feature extraction.

We have so far applied learning metrics to two widely used unsupervised information visualization methods: the Self-Organizing Map and Sammon's mapping, a sample Multidimensional Scaling (MDS) method.

Computation of approximations to the metric

Globally the learning metric (11.1) becomes minimal path integrals of local distances. The local distances in turn are based on conditional auxiliary densities $p(c|\mathbf{x})$. For practical computation, the densities must be estimated and the minimal path integrals approximated. We have developed several approximations; the choice needs a tradeoff between computation time and accuracy.

Several semiparametric estimators of the conditional density $p(c|\mathbf{x})$ are available. The still open theoretical question is how to choose the estimator rigorously.

The simple approximation for the distance between two points $\mathbf{x}_1$ and $\mathbf{x}_2$ is the local quadratic form [2]

$$d_1^2(\mathbf{x}_1, \mathbf{x}_2) = (\mathbf{x}_2 - \mathbf{x}_1)^T \mathbf{J}(\mathbf{x}_1)(\mathbf{x}_2 - \mathbf{x}_1) \tag{11.2}$$

called the 1-point approximation. An improved version called the T-point approximation [3] computes the metric at T points between the start and end point, yielding

$$d_T(\mathbf{x}_1, \mathbf{x}_2) = \frac{1}{T} \sum_{i=0}^{T-1} \left(\mathbf{r}^T \mathbf{J} \left(\mathbf{x} + \frac{i}{T} \mathbf{r} \right) \mathbf{r} \right)^{1/2} , \tag{11.3}$$

where $\mathbf{r} = \mathbf{x}_2 - \mathbf{x}_1$.

Both approximations assume the minimal path is a line. A further improvement is to form a graph whose edge weights are pairwise T-point distances between data points and perform a graph search for the minimal path [4]. This is called the graph approximation; it allows both linear and piecewise linear paths. Since data points are used as graph vertices, distances are computed more accurately where the data is dense.

Information visualization methods

The sequential SOM algorithm iterates *winner selection* and *adaptation*. In the learning metric the winner is sought by

$$w(\mathbf{x}(t)) = \arg\min_i d_L^2(\mathbf{x}(t), \mathbf{m}_i(t)) . \tag{11.4}$$

where t is the iteration, $\mathbf{x}(t)$ is the input and d_L can be either the local distance approximation d_1 or the T-point approximation d_T. The latter is more accurate, but computationally heavier.

For the local approximation the adaptation step can be shown to equal the familiar SOM learning rule,

$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) + \alpha(t) h_{wi}(t)(\mathbf{x}(t) - \mathbf{m}_i(t)) , \tag{11.5}$$

were $\alpha(t)$ is the learning rate and $h_{wi}(t)$ is the neighbourhood function.

In empirical tests the SOM-L with the improved (T-point) distance approximation significantly outperforms the 1-point SOM-L as well as classical SOM and a supervised SOM [3, 4].

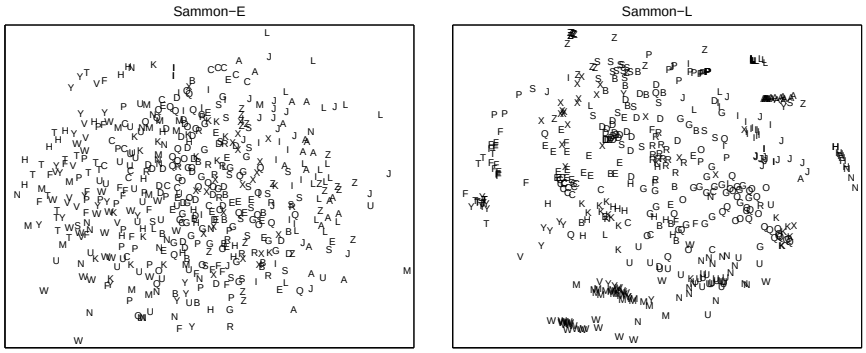


Figure 11.1: Sammon's mapping in learning metrics (right) separates the different letters of the Letter Recognition data (from UCI Machine Learning Repository) set clearly better than the Sammon's mapping in the Euclidean metric (left).

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

Sammon's mapping, as well as the other MDS methods, are based on the pairwise distance matrix, are ideal candidates for the graph approximation since they are based on the pairwise distance matrix. The distances need to be computed only once.

The difference to the traditional Sammon's mapping where the pairwise distance matrix is computed in the Euclidean metric is illustrated in Figure 11.1 [4]. The class separation is clearly increased when the learning metrics is used, but the topology of the samples is still retained.

11.3 Discriminative Clustering (DC)

The original motivation for discriminative clustering was its asymptotic equivalence to vector quantization in learning metrics. DC turned out to have other interesting interpretations as well: It extends earlier works on mutual information maximization (IMAX [5], Information Bottleneck [6]) and connects learning metrics to generative models and contingency tables.

DC partitions a vectorial data space to a set of connected partitions that are homogeneous by distributions of an auxiliary variable or variables present in the data [7]. The homogeneity criterion of partitions turns out to be equivalent to informativeness of the partitions of the auxiliary variable(s). Membership of a sample in a partition then tends to predict the value of the auxiliary variable well, and vice versa. Still, the partitions are solely defined in terms of the primary data, without reference to the values of auxiliary data. Hence future data without the associated auxiliary variable can be partitioned. The relative locality of the clusters in the primary data space makes them useful for exploratory analysis.

A prototypical application would be segmenting customers of a company in terms of background information, but by using buying behaviour as the criterion of segment homogeneity. Buying behaviour guides the segmentation but does not directly define the segments. Incoming customers without buying history can then be immediately assigned to the predefined segments. Other applications include, e.g., understanding company bankruptcy, finding relationships between gene expression databases (Section 10.3), and guiding text document clustering with classifications of informaticians.

For densities. The original formulation for DC is for probability densities $p(c, \mathbf{x})$ of auxiliary data c and primary data $\mathbf{x}$. This version is easy to understand, but directly applicable only for large data/cluster rations.

Partitions of the primary data are restricted to be Voronoi regions, which makes them connected, relatively local, and therefore easy to interpret. Homogeneity of the auxiliary data distributions within the clusters is measured by the intra-cluster Kullback-Leibler divergence

$$E = \int D_{\mathrm{KL}}(p(c|\mathbf{x}) \| \boldsymbol{\psi}_{j(\mathbf{x})})\, p(\mathbf{x})\, d\mathbf{x} \,, \tag{11.6}$$

which is minimized with respect to the distributions of auxiliary data within clusters, $p(c|\mathit{Cluster}_j) \equiv \boldsymbol{\psi}_j$ and the Voronoi partitioning defined by the centroid parameters $\mathbf{m}_j$ (implicit in assignments $j(\mathbf{x})$). Minimizing the distortion is equivalent to maximizing the informativeness of the clusters about the values of the auxiliary variable, in the sense of mutual information. Gradient algorithms can be applied if the partitions are first smoothed. An extremely simple on-line learning rule results.

For data sets. The log-likelihood of a piece-wise constant model for the conditional densities $p(c|\mathbf{x})$ approaches the distortion (11.6) when the size of the data set grows. It is therefore a good candidate for the cost function of DC for finite data sets [8]. From the viewpoint of clustering, the distributional prototypes $\boldsymbol{\psi}$ are not interesting and can then be marginalized out, which leads to a likelihood only depending on the Voronoi partitioning $\{\mathbf{m}_j\}$:

$$L_{\mathrm{DC}}(\{\mathbf{m}_j\}) \propto \sum_{ij} \log \Gamma(n_i^0 + n_{ji}) - \sum_j \log \Gamma(N^0 + N_j) \,, \tag{11.7}$$

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 $N_j = \sum_i n_{ji}$, $N^0 = \sum_i n_i^0$.

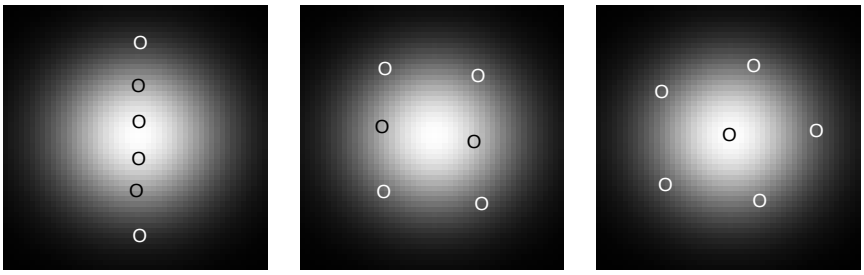


Figure 11.2: Discriminative clustering of simple toy data, where only the vertical direction is indicated to be relevant by auxiliary data associated to the 2D primary data. The primary data is sampled from the Gaussian distribution (grey shades), while the conditional distribution of the auxiliary data changes in the vertical direction. The regularized solution (middle) shares properties of the discriminative solution and the K-means solution (right).

After the partitions are smoothed the new cost function can be optimized by gradient algorithms. Direct optimization by simulated annealing is also possible, but a simple conjugate gradient algorithm with smoothed partitions leads to equally good results and is faster. In empirical comparisons the marginaled finite-data model has been found to outperform the simple on-line algorithm resulting from (11.6).

Regularization. Tests indicate that the performance of the purely discriminative DC algorithms is improved if the cost functions are 'regularized' by partially taking into account the margin distribution $p(\mathbf{x})$ in one way or another (Fig. 11.2). Note that taking it fully into account would lead to modeling the joint distribution $p(c, \mathbf{x})$, which is different by its goal and empirically shown to be inferior in the task of DC.

Non-Euclidean spaces. DC has been extended for data on hyperspheres and on distributional spaces. The latter formulation is applicable to text documents under the usual 'bags of words' assumption, where word frequencies are analyzed and the order of words in the documents is ignored. The method has been applied to scientific texts from the INSPEC database [9], by using keywords chosen by the document authors as auxiliary data. Keywords improve feature selection in the full documents and therefore improve clustering results compared to classic methods.

Connection to learning metrics. For a large number of clusters, DC performs vector quantization in learning metrics (Section 11.1): The Euclidean distortion of normal vector quantization becomes replaced with a distortion computed in the Fisher metric (11.1). The Fisher metric measures changes in the conditional distributions $p(c|\mathbf{x})$ of the auxiliary variable [10].

The asymptotic connection was utilized in practice by plugging a local approximation of Fisher metrics to standard K-means clustering [11]. The adaptable metric frees the Voronoi partitions from being defined in Euclidean metric and allows more optimal shapes. In tests the resulting algorithms, although computationally heavy, have outperformed the plain DC.

Figure 11.3: The Helsinki capital area segmented into Voronoi regions maximally informative of demographics. Associative clustering of geographic coordinates and vectorial sosiodemographic data finds segments for both 'margin spaces'. In the figure, only one margin space, the geography, is shown. Demographically distinct and homogeneous regions such as the downtown become clearly separated. Similar clusters become defined to the high-dimensional sosiodemographic space.

Associative clustering: bidirectional DC

Discriminative clustering quantizes a continuous variable and then maximizes statistical dependency between two discrete variables: the partitions and the auxiliary variable guiding the partitioning. Contingency tables are a classic framework for quantifying and testing such dependencies. In this framework, the cost (11.7) is interpretable as a *Bayes factor* between the hypotheses of dependent and independent margins [12].

In DC one margin is fixed. We have called the generalization to two adjustable margins *associative clustering* (AC; [13]). Then two vectorial variables are quantized by Voronoi partitionings, and the partitionings are adjusted to maximize their mutual dependency in the sense of the Bayes factor. Techniques similar to discriminative clustering can be applied, including the regularization methods and smoothed partitions. A demonstration of AC is shown in Figure 11.3.

11.4 Discriminative components

Unsupervised principal components and factor analyses search for components of data that can be used for data exploration, visualization and dimensionality reduction.

A classical method for supervising the components is linear discriminant analysis (LDA). It has been commonly used for two tasks: the more common one is linear classification (supervised learning), but the components can also be used for exploring and visualizing class differences. We have generalized LDA for this latter purpose, but searching for linear components that are more generally *informative of* or *relevant to* the the classes of samples. The task of extracting components relevant to auxiliary data could perhaps be called Relevant Component Analysis.

We search for linear relevant components [14] by optimizing the linear projection $\mathbf{y} = \mathbf{f}(\mathbf{x}) = \mathbf{W}^T\mathbf{x}$, where the columns of $\mathbf{W}$ are the component directions. The criterion is simply maximization of the log-likelihood of the auxiliary data given the projection, i.e.,

$$L = \sum_{(\mathbf{x},c)} \log \hat{p}(c|\mathbf{f}(\mathbf{x})) \tag{11.8}$$

where c are the auxiliary data and $\hat{p}$ is an estimator computed after the projection.

The key point in this method is its simplicity. The likelihood is a well-defined, simple criterion for fitting a projection to finite data, yet it has interesting theoretical connections and works better than alternative methods in practice. Maximizing the likelihood is asymptotically equivalent to maximizing the mutual information $I(C, \mathbf{f}(X))$ when consistent estimators $\hat{p}$ are used. Moreover, maximizing the likelihood is asymptotically approximately equivalent to minimizing a reconstruction error in learning metrics under some assumptions, so the components can be considered principal components in learning metrics.

The method has empirically outperformed classical and recent [16] methods. It has been applied to bioinformatics (Chapter 10) and assessing convergence of MCMC simulations (below).

11.5 Visualization of posterior distributions

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 [17] that the main multivariate convergence measure, the multivariate potential scale reduction factor (MPSRF) developed by Brooks and Gelman [18], equals the cost function of a one-dimensional linear discriminant analysis (LDA), a method that discriminates between data classes.

MCMC chains have traditionally been visualized by time series plots, marginal histograms or 2-dimensional scatter plots of two variables. The problem with these visualizations is that they 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 reduce the number of visualizations. A scatter plot of a projection on the two best discriminative components (see Figure 11.4) is the single best two-dimensional image in the sense of the MPSRF measure.

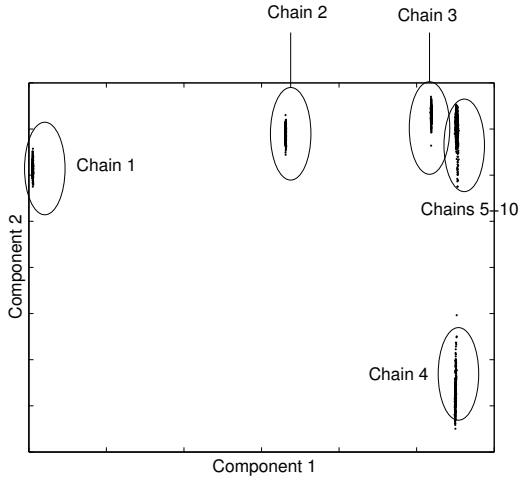


Figure 11.4: Two-dimensional LDA projection of samples from a MCMC simulation that does not converge. Chains 1-4 have gotten stuck in a degenerate state. The ellipses have been drawn by hand to mark the chains.

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 the generalization of LDA introduced in Section 11.4.

Sometimes we are interested in visualizing the posterior distribution for other reasons than studying convergence of a sampler. We might for example be interested how the parameters affect the model output. Toward this end, we have proposed [19] a method that uses the Fisher metric of the model with a non-linear projection method, to create visualizations of the posterior that reflect the effect parameters have on the output.

References

[1] Shun-ichi Amari and Hiroshi Nagaoka. *Methods of Information Geometry*, volume 191. American Mathematical Society and Oxford University Press, 2000.

[2] Samuel Kaski, Janne Sinkkonen, and Jaakko Peltonen. Bankruptcy analysis with self-organizing maps in learning metrics. *IEEE Transactions on Neural Networks*, 12:936–947, 2001.

[3] Jaakko Peltonen, Arto Klami, and Samuel Kaski. Learning more accurate metrics for self-organizing maps. In José R. Dorronsoro, editor, *Artificial Neural Networks—ICANN 2002*, pages 999–1004. Springer, Berlin, 2002.

[4] Jaakko Peltonen, Arto Klami, and Samuel Kaski. Learning metrics for information visualization. In *Proceedings of the Workshop on Self-Organizing Maps (WSOM'03)*, pages 213–218. Hibikino, Kitakyushy, Japan, September 2003.

[5] Suzanna Becker. Mutual information maximization: models of cortical self-organization. *Network: Computation in Neural Systems*, 7:7–31, 1996.

[6] Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. In *37th Annual Allerton Conference on Communication, Control, and Computing*, pages 368–377. Urbana, Illinois, 1999.

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

[8] Janne Sinkkonen, Samuel Kaski, and Janne Nikkilä. Discriminative clustering: Optimal contingency tables by learning metrics. In T. Elomaa, H. Mannila, and H. Toivonen, editors, *Proceedings of the ECML'02, 13th European Conference on Machine Learning*, pages 418–430. Springer, Berlin, 2002.

[9] Jaakko Peltonen, Janne Sinkkonen, and Samuel Kaski. Discriminative clustering of text documents. In Lipo Wang, Jagath C. Rajapakse, Kunihiko Fukushima, Soo-Young Lee, and Xin Yao, editors, *Proceedings of ICONIP'02, 9th International Conference on Neural Information Processing*, pages 1956–1960. IEEE, Piscataway, NJ, 2002.

[10] Samuel Kaski and Janne Sinkkonen. Principle of learning metrics for data analysis. *The Journal of VLSI Signal Processing-Systems for Signal, Image, and Video Technology, Special Issue on Data Mining and Biomedical Applications of Neural Networks*, accepted for publication.

[11] Jarkko Salojärvi, Samuel Kaski, and Janne Sinkkonen. Discriminative clustering in Fisher metrics. In O. Kaynak, E. Alpaydin, E. Oja, and L. Xu, editors, *Artificial Neural Networks and Neural Information Processing - Supplementary proceedings ICANN/ICONIP 2003*, pages 161–164. Istanbul, Turkey, June 2003. To appear.

[12] I. J. Good. On the application of symmetric Dirichlet distributions and their mixtures to contingency tables. *Annals of Statistics*, 4(6):1159–1189, 1976.

[13] J. Sinkkonen, J. Nikkilä, L. Lahti, and S. Kaski. Associative clustering by maximizing a bayes factor. Technical Report A68, Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, Finland, 2003.

[14] 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.

[15] Kari Torkkola and William Campbell. Mutual information in learning feature transformations. In *Proceedings of the 17th International Conference on Machine Learning*, pages 1015–1022. Morgan Kaufmann, Stanford, CA, 2000.

[16] Kari Torkkola. Feature extraction by non-parametric mutual information maximization. *Journal of Machine Learning Research*, 3:1415–1438, 2003.

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

[18] 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.

[19] Jarkko Venna and Samuel Kaski. Visualizing high-dimensional posterior distributions in bayesian modeling. In O. Kaynak, E. Alpaydin, E. Oja, and L. Xu, editors, *Artificial Neural Networks and Neural Information Processing - Supplementary proceedings ICANN/ICONIP 2003*, pages 165–168, Istanbul, Turkey, June 2003.

Chapter 12

Natural language processing

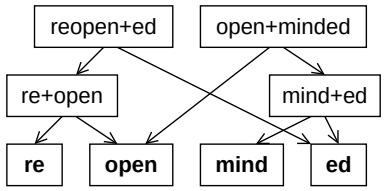
Krista Lagus, Mathias Creutz, Mikko Kurimo, Krister Lindén

12.1 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 would be 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 is infeasible, as the number of possible word forms is very high. The productivity of word forming in Finnish is illustrated in Figure 12.1a, where one single word consists of six morphemes.

Morph.	kahvi	n	juo	ja	lle	kin
Transl.	coffee	of	drink	-er	for	also

(a)



(b)

Figure 12.1: (a) Morphological segmentation of the Finnish word for "also for [the] coffee drinker". (b) Hypothetical binary splitting trees for the words "reopened" and "open-minded" (segmented as "re+open+ed" and "open+mind+ed").

We have developed language-independent, data-driven methods for the unsupervised segmentation of the words in a corpus. We call the resulting segments *morphs* and we do not distinguish between categories, such as stems, suffixes, and prefixes. For us words simply consist of (possibly lengthy) sequences of morphs concatenated together. In this sense, our work differs from most previous work in the field of automated morphology learning, where more limitations are set on word structure (e.g., [1]). Instead, our work resembles algorithms for unsupervised text segmentation and word discovery (e.g., [2]).

In [3] we present a model inspired by the Minimum Description Length (MDL) principle. This means that we try to find the optimal balance between *accuracy* of representation and model *complexity*, which generally improves generalization capacity by inhibiting overlearning. In concrete terms, we construct a morph vocabulary, or a *lexicon of morphs*, so that it is possible to form any word in the corpus by the concatenation of some morphs. Each word in the corpus is rewritten as a *sequence of morph pointers*, which point to entries in the lexicon. We aim at finding the optimal lexicon and segmentation, i.e., a set of morphs that is concise, and moreover gives a concise representation for the corpus.

The optimal segmentation is obtained by splitting words recursively and trying to find common subword chunks, which are potential morphs. Figure 12.1b shows a hypothetical recursive splitting of two English words. The leaf nodes correspond to morphs discovered by the algorithm.

Results

For evaluation purposes, we have constructed a gold standard segmentation based on linguistic theory for 1.4 million Finnish and twenty thousand English word forms. When comparing our MDL-inspired method to the main other method, called Linguistica (cf. [1]), we obtain clearly better results on Finnish, and for very small data sets on English, whereas Linguistica is better on English for larger data sets. However, Linguistica utilizes some linguistic assumptions particularly suitable for Indo-European languages.

aamuyö + stä, elma + n, hyvinvointi + yhteis + kuntina, kellari + ssa + kaan, lait + tomasti + kin, miljonääri + ltä, palkka + tuloilla + an, rebrov + ia, sunnuntai + vuorossa, tuuli + potentiaali + sta, wal + ston + in, ääni + lehtenä
abandon, a + shore, book + ers, cherry, cooper + s, dinner + s, entrance, form + ing, harpsichord + ist, in + jury, learned, men + a + c + ing, n + un, pick + ers, radio + activity, sir, succeed + ing, travel + ler, war + 's

Figure 12.2: Examples of Finnish and English words segmented by our algorithm.

Some sample segmentations are shown in Figure 12.2. They include correctly segmented words, where each boundary coincides with a real morpheme boundary (e.g., "kellari+ssa+kaan", "miljonääri+ltä", "dinner+s", "form+ing", "war+'s"). In addition, there are over-segmented words, with boundaries inserted at incorrect locations (e.g., "men+a+c+ing" instead of "menac+ing"), as well as under-segmented words, where some boundary is missing (e.g., "learned" instead of "learn+ed"; "ääni+lehtenä" instead of "ääni+lehte+nä"). Sometimes many alternative segmentations seem correct: e.g., "hyvin+vointi", "hyv+in+voi+nti".

In [4] we re-formulated the model in a probabilistic framework and studied whether prior information about morph length and frequency could be utilized to avoid over- and under-segmentation. This did not lead to considerable improvements in overall accuracy, though. So far each morph has been considered individually, irrespective of the previous and next morph. In future research, the model will be extended to learn information on morph categories and sequences of them.

Demonstration and applications

An online demonstration of the model is available at the address: `http://www.cis.hut.fi/projects/morpho/`. The demonstration allows the user to select a corpus to be used as training data and to type in words that are then segmented according to the model.

The algorithm in [3] has been applied for producing morph vocabularies for *automatic speech recognition*, both for Finnish [5] and Turkish [6] (cf. Section 13.2). Among the different vocabulary approaches tested, the ones based on morphs were most successful.

Acknowledgement

We appreciate the contribution of Sami Virpioja, who implemented the online demo.

References

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

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

[3] M. Creutz and K. Lagus. Unsupervised discovery of morphemes. In *Proc. Workshop on Morphological and Phonological Learning of ACL'02*, pages 21–30, Philadelphia, 2002.

[4] M. Creutz. Unsupervised segmentation of words using prior distributions of morph length and frequency. In *Proc. ACL'03*, pages 280–287, Sapporo, 2003.

[5] V. Siivola, T. Hirsimäki, M. Creutz, and M. Kurimo. Unlimited vocabulary speech recognition based on morphs discovered in an unsupervised manner. In *Proc. Eurospeech'03*, pages 2293–2296, Geneva, 2003.

[6] 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, 2003.

12.2 Word sense disambiguation using document maps

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 [1]. In particular, Yarowsky [2] noted that words tend to keep the same sense during a discourse.

In [3] 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 Section 8.4 or [4]). 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 12.3.

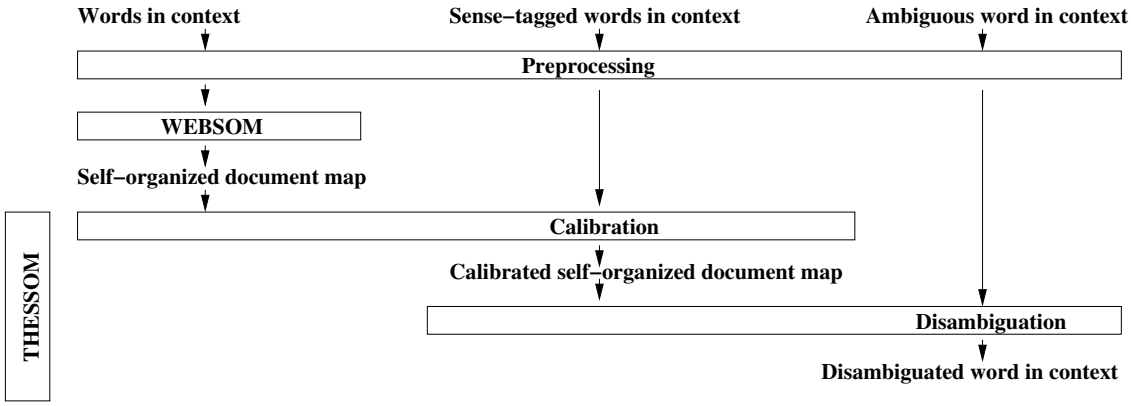


Figure 12.3: Data flow of word sense disambiguation with self-organized document maps

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 [5], 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.

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] C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts, 1999.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

12.3 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 13.2) 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.

It has been previously shown that good topically organized clustering of large text collections can be achieved efficiently using the WEBSOM method (see Section 8.4 or [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 12.4.

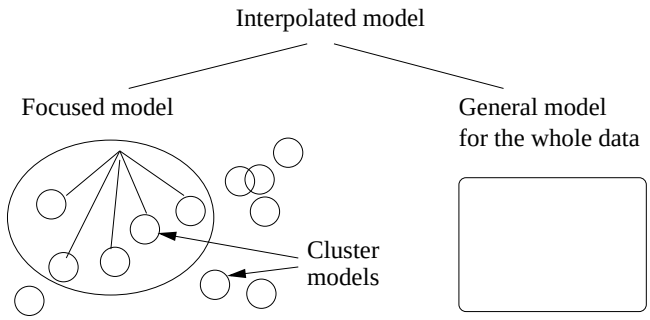


Figure 12.4: *A focusing language model obtained as an interpolation between topical cluster models and a general model.*

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.

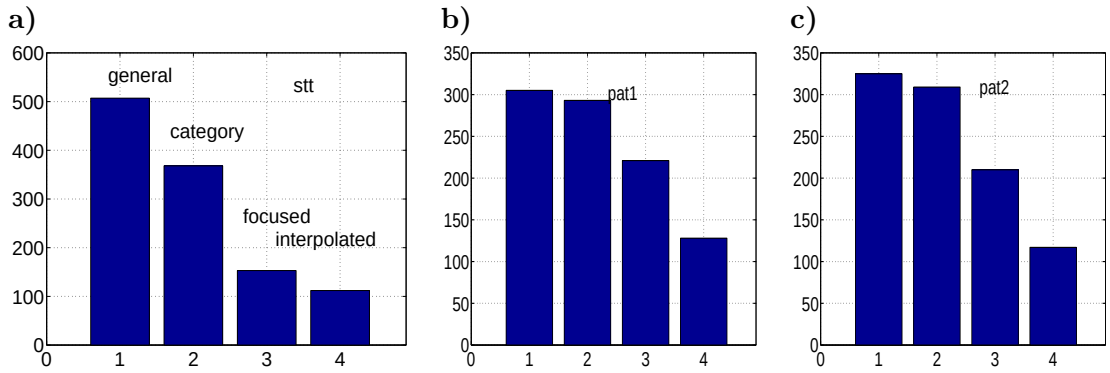


Figure 12.5: The perplexities of the different language models, **a)** for the Finnish STT news corpus, **b)** for smaller patent corpus and **c)** for larger patent corpus. The explanation of the bars in each figure, from left to right: 1. general model for the whole corpus, 2. category-specific model 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 12.5 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.

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.

12.4 Semantic analysis of Finnish words and sentences

Observation of language use provides indirect evidence of the representations that humans utilize. The study of conceptual, 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.

As shown in [1,2], the self-organizing map (Chapter 8) can be applied for clustering English word forms based on the words that have appeared in their immediate contexts. In Finnish, however, the rich inflectional morphology poses a challenge as the vocabularies built on inflected word forms are typically very large. Moreover, also the inflections, some of which correspond to prepositions and function words in English, carry relevant semantic information [3]. Furthermore, the much less restricted word order compared to English is likely to cause more variation in the immediately nearby words.

In this project, we have applied the SOM algorithm to visualize and cluster common Finnish verbs based on averaged contextual features. The verb category was selected for study for two reasons: there exists a semantic reference classification of Finnish verbs [3] for comparing the results, and the semantic representation of verbs is considered an interesting problem in linguistics [3,4] because of the richness and variability of information that is connected to different verbs. In collecting information about the verbs, both morphosyntactic properties (inflections) [5] and noun base forms [6] were examined, and the resulting categorizations were compared.

An example of a map where 600 Finnish verbs were organized based on their contextual morphological features is shown in Figure 12.6. In contrast, the use of nouns as features produced for example the kinds of verb categories shown in Table 12.1.

Table 12.1: Sample verb categories based on noun categories.

Finnish verbs	Translations	Topic
myydä, ostaa, tuottaa, palkata, työllistää, kattaa, vuokrata	sell, buy, produce, hire, employ, cover, rent	business
nousta, laskea, kasvaa, pudota, vähentyä, kohota, pienentyä, supistua, noutaa, kallistua	rise, decrease, grow, fall, diminish, rise, get smaller, contract, fetch, go up in price	stock rates
kuolla, hukkua, ampua, surmata, hyökätä, loukkaantua, menehtyä	die, drown, shoot, kill, attack, get hurt, pass away	dying

The results in all the experiments show that even the simple contextual features used, collected over a large number of instances, can be suitable for obtaining automatically a semantic clustering and organization of verbs. In general, morphosyntactic properties seem to push the categorization towards the direction of linguistic semantics, while categorization based on nouns or noun categories is more a reflection of the topics of their

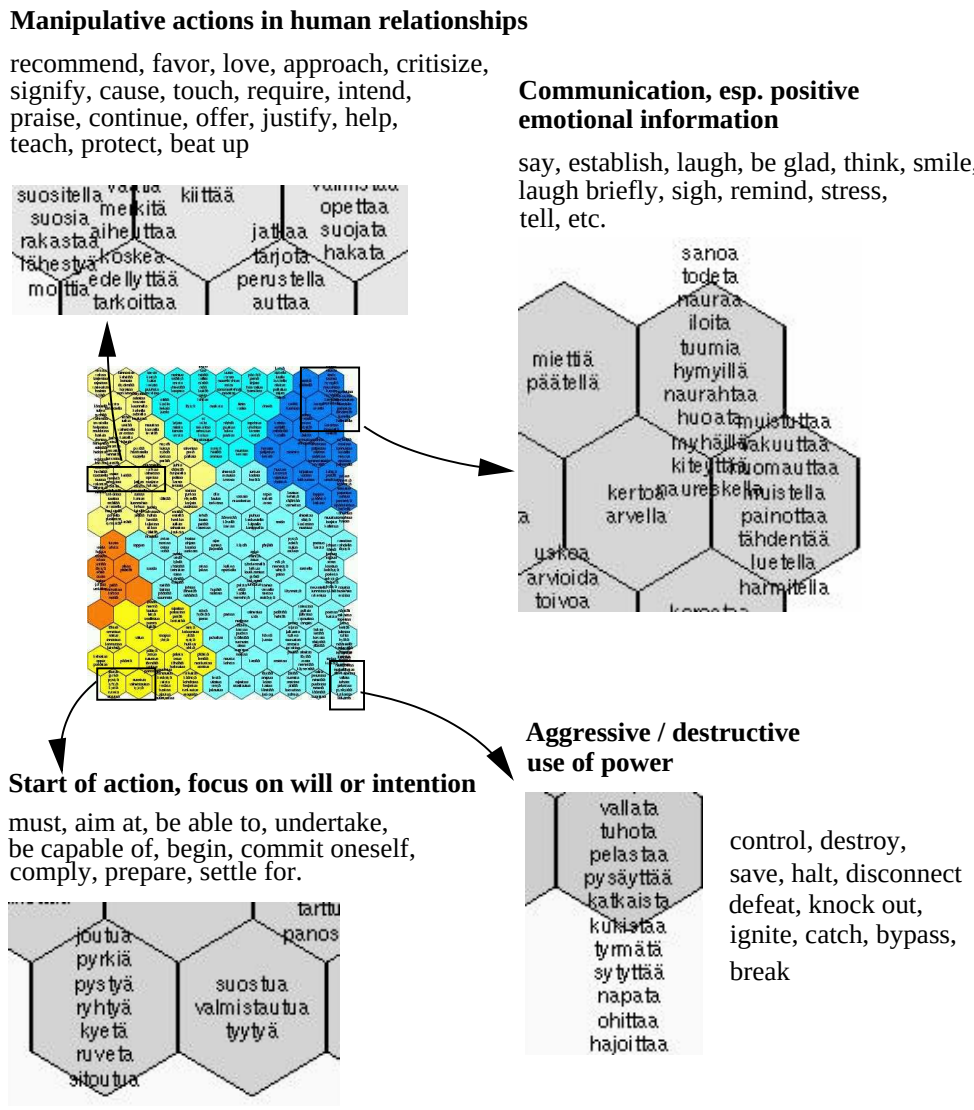


Figure 12.6: A map of the 600 most frequent verbs (base forms) in the Newspaper corpus. The verbs were organized on the basis of the distribution of morphological features in one preceding and two succeeding words, collected over all instances of the verb in any inflected form. The contents of four sample map regions are shown in the insets. In the reference classification (pp. 157–165 in [3], many of the verbs e.g. in the lower right corner indicating 'destructive use of power' are further divided into two specific categories, namely (1) break verbs (*tuhota* 'destroy', *katkaista* 'break', *hajoittaa* 'break down') and (2) fight verbs (*pysäyttää* 'stop', *kukistaa* 'defeat', *tyrmätä* 'knock out'). Similar categories can be found in [4] for English verbs.

corresponding texts. When compared to a reference classification of Finnish verbs [3], clustering shows a somewhat different perspective or world view than Pajunen's. In particular, the organization of verbs on the map reflects the importance of cultural, social, and emotional dimensions in lexical organization.

References

[1] T. Honkela. *Self-Organizing Maps in Natural Language Processing.* PhD thesis, Helsinki University of Technology, 1997, Espoo, Finland.

[2] H. Ritter and T. Kohonen. Self-organizing semantic maps. *Biological Cybernetics*, 1989; 61:241-254

[3] A. Pajunen. *Argumenttirakenne. Asiaintilojen luokitus ja verbien käyttäytyminen suomen kielessä.* Suomalaisen Kirjallisuuden Seura, 2001.

[4] B. Levin. *English Verb Classes and Alternations: a Preliminary Investigation.* The University of Chicago Press, Chicago and London, 1993.

[5] K. Lagus and A. Airola. Analysis of functional similarities of Finnish verbs using the self-organizing map. In *ESSLLI'01 Workshop on The Acquisition and Representation of Word Meaning,* August 2001.

[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.

Chapter 13

Speech recognition

Mikko Kurimo, Panu Somervuo, Vesa Siivola, Teemu Hirsimäki

13.1 Acoustic modeling

Acoustic modeling in automatic speech recognition (ASR) means building statistical models for subword units based on the feature vectors computed from speech. Feature representation is an important part of any pattern recognition system and ASR is no exception. It is difficult to develop any theoretically optimal feature extraction method which would minimize the recognition error. Usually the discriminative training is applied to the estimation of the model parameters and the feature representation is more or less fixed, see e.g. [1]. In practice, several feature extraction methods have been experimented and during the long history of ASR, some of them have been experimentally proved to be more beneficial than others.

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. Computation of delta features can be considered as a fixed transformation to the block of original feature vectors. We have experimented also other linear and nonlinear feature transformations.

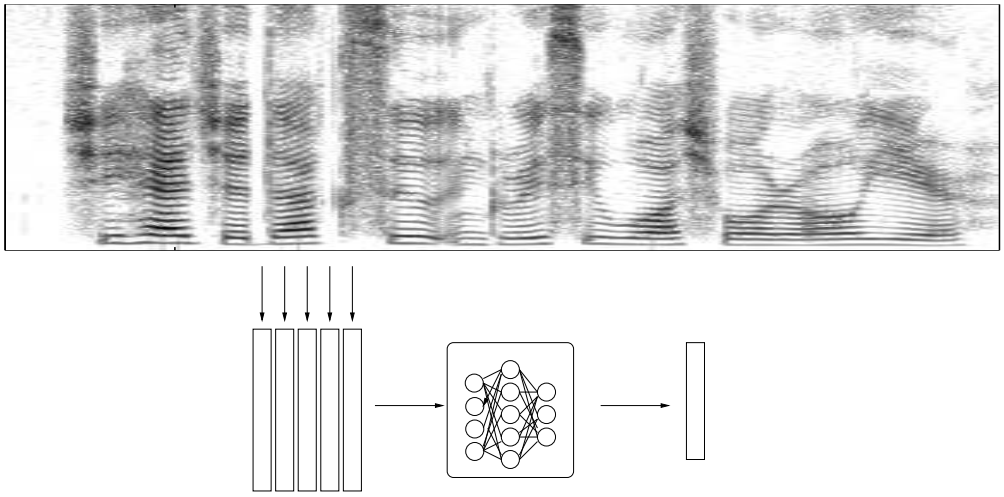


Figure 13.1: Feature transformation. One or more frames (five in this figure) original feature vectors, e.g. logarithmic mel-spectra are fed to the linear (matrix) or nonlinear (MLP network) feature transform which performs the projection of the original feature vector (or concatenation of them) to the new feature space. The output is used as a feature vector in the mixture-of-Gaussians based HMM system.

In our experiments, unsupervised transformations were based on principal component analysis (PCA) and independent component analysis (ICA) and discriminative transformations were based on linear discriminant analysis (LDA) and multilayer perceptron (MLP) networks. These transformations were experimented in TIMIT phone recognition [2] where clear improvements were gained in the recognition rate compared to the baseline MFCC feature. In another experiment [3], the acoustic models were trained using 60 hours of HUB5 training data and they were tested using OGI Numbers corpus. The combination of the PLP cepstrum and the MLP network based feature transformation stream gave the best result. The baseline word error rate was reduced from 4.1 % to 3.1 %.

Currently used speech recognizers are typically based on hidden Markov models where HMM states are modeled by Gaussian mixtures. In order to avoid the large number of parameters in the model, the covariance matrices of the Gaussians are diagonal. We have experimented the maximum likelihood linear transformation (MLLT) [4], which takes the diagonal Gaussian assumption into account when forming the transformation. The result is not the global PCA transform, since in our case the data is not modeled by a single Gaussian with a single covariance matrix but each speech unit is modeled by its own mixture of Gaussians where the diagonal covariance matrices need not be the same. Using the MLLT framework, feature transformations based on heteroscedastic linear discriminant analysis (HLDA) can also be constructed. Contrasted to the basic LDA, HLDA does not assume equal class covariance matrices. Applying these transformations to Finnish speech recognition system gave very promising results:

	monophone HMMs		triphone HMMs	
	letter error %	word error %	letter error %	word error %
baseline MFCC	11.0	44.0	4.7	24.8
MFCC+MLLT	9.0	40.2	4.5	24.1
MFCC+HLDA	8.4	37.5		

Besides speech recognition, we have also investigated methods for representing high-dimensional feature vectors. In [5] it was studied how to capture the intrinsic dimensionality of speech using fractal-dimensionality measure, multi-dimensional scaling, and hypercubical Self-Organizing Map. These results can give insights to the data being modeled and that way contribute also to the developments in speech recognition.

State-of-the-art speech recognizers are complex systems with large number of parameters. This raises the challenge how to get robust estimates and what is the optimal number of model parameters. One elegant way is to use Bayesian modeling. In [6], standard maximum likelihood (ML) estimation was compared to the variational Bayesian approach for training mixtures of Gaussians. Advantages of Bayesian approach were clear: estimation converged faster, there was no tendency of overfitting, and the likelihoods of unseen test data were better for any given number of mixture components.

References

[1] P. Somervuo. Two-level phoneme recognition based on successive use of monophone and diphone models. In *Proc. EUSIPCO*, vol. 3, pages 77–80, 2002.

[2] P. Somervuo. Experiments with linear and nonlinear feature transformations in HMM based phone recognition. In *Proc. IEEE ICASSP*, vol. 1, pages 52–55, 2003.

[3] P. Somervuo, B. Chen, and Q. Zhu. Feature transformations and combinations for improving ASR performance. In *Proc. Eurospeech*, vol. 1, pages 477–480, 2003.

[4] M. Gales. Semi-tied covariance matrices for hidden Markov models. *IEEE Tr. SAP*, 7(3), pages 272–281, 1999.

[5] P. Somervuo. Speech dimensionality analysis on hypercubical self-organizing maps. *Neural Processing Letters*, 17(2), pages 125–136, 2003.

[6] P. Somervuo. Speech modeling using variational Bayesian mixture of Gaussians. In *Proc. ICSLP*, pages 1245–1248, 2002.

13.2 Language modeling

Language model unit selection for speech recognition

The traditional method to model language for speech recognition is the n-gram model. The probability of a new word is estimated based on a few previous words. For Finnish, estimating the n-gram probabilities is difficult, since there is a vast number of different word forms. For example, a single verb has theoretically thousands of inflected word forms. We have chosen to split the words to smaller units to have fewer probabilities to estimate and to cover a larger vocabulary. As subword units we have evaluated syllables and statistically found morpheme-like units [1].

For Finnish, words can be split into syllables based on a few simple hyphenation rules, except on boundary between parts of a compound word. Our algorithm implements the simple ruleset and makes infrequently mistakes on compound words. A morpheme is the smallest meaning bearing element of a word. We have used an automatic statistical method for finding morpheme-like units, called morphs (see Section 12.1).

In our evaluations, using syllables for language model units decreases the recognition word error rate 22% relative to word based model. Using morphs reduces the word error rate 44% relative to word based model. The morphs seem to be better suited for language modeling, since each morph has a distinct meaning which is useful for language modeling. For syllable and morph based models, we have another advantage: we do not need to know all of the words of Finnish language, since the words can be constructed from the smaller pieces.

To assess the language-independence of the word splitting method, we applied the same algorithm to Turkish, which is another agglutinative language[1]. To compare the performance with baseline speech recognizer, the n-gram models were trained both to these new data-driven and old rule-based morphemes and words. The data-driven morphemes achieved clearly the lowest error rates in all large-vocabulary continuous speech recognition tests made [2]. The work with Turkish data is done in a close collaboration with the University of Colorado in Boulder and the Middle East Technical University.

Focusing language models in speech recognition

The efficient language processing tools developed in the laboratory (WEBSOM) have been applied to organize language models based on the topical structure of the discourse [3]. The objective is to increase the language modeling accuracy and to obtain improved speech recognition results by automatically detecting and focusing into the best available language model for the recognition task at hand. This work is done in a close collaboration with the Natural language modeling group (see Section 12.3).

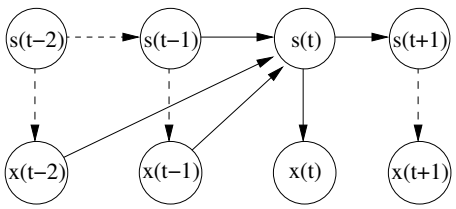


Figure 13.2: The state-space model for language modeling. $\mathbf{s}(t)$ is the current state and $\mathbf{x}(t-1)$ the previous observation.

[1] In agglutinative languges words frequently have multiple suffices concatenated one after each other.

State-space method for language modeling

The most common language model for speech recognition is the n-gram model. With back-off and smoothing, it provides a relatively robust model.However, the n-gram model cannot generalize from similar words: seeing a phrase like "Monday morning was clear" does not help modeling the phrase "Tuesday evening is cloudy" at all. This kind of generalization can be achieved by clustering similar words together and interpolating this cluster n-gram with a regular n-gram.

We have tried to achieve the generalization by mapping the words to n-dimensional feature space, so that similar words are mapped close to each other. The probability of a word is calculated as a smooth function of the features and the previous state, leading to good generalization. This kind of approach with neural networks has been shown to yield good results [4]. The mathematics of our method are based on linear state-space modeling, which is also used in famous algorithms like Kalman filtering. We have added explicit dependencies to previous observations to make the teaching of the model simpler (see Fig. 13.2).

During the first experiments, we simply tried predicting the next letter based on previously seen letters, since the learning algorithm was computationally extremely demanding [5]. We are currently working on making the algorithm suitably fast for word prediction. Figure 13.3 shows a hypothetical idealized picture of both the feature and the internal state of the model.

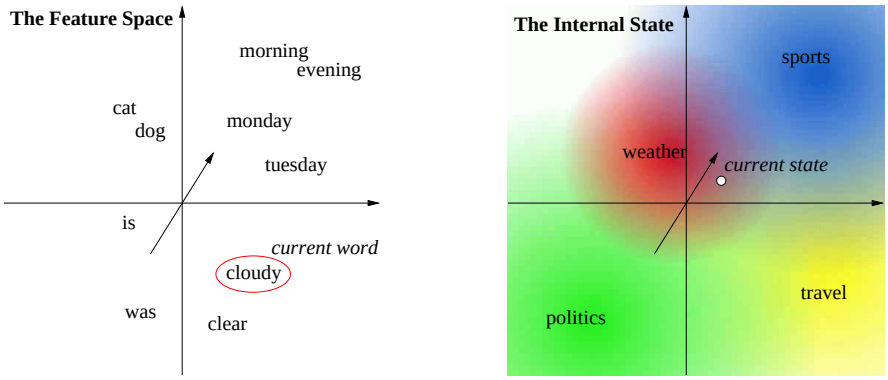


Figure 13.3: The ideal state-space language model. On left is the feature space and on right the internal state of the model.

References

[1] V. Siivola, T. Hirsimäki, M. Creutz and M. Kurimo: Unlimited vocabulary speech recognition based on morphs discovered in an unsupervised manner, *Eurospeech03*, pages 2293-2296, 2003.

[2] K. Hacioglu, B. Pellom, T. Ciloglu, O. Ozturk, M. Kurimo, and M. Creutz. On Lexicon Creation for Turkish LVCSR. *Eurospeech03*, pages 1165-1168, 2003.

[3] M. Kurimo and K. Lagus. An efficiently focusing large vocabulary language model. *ICANN'02*, pages 1068–1073, Madrid, Spain, 2002.

[4] Y. Bengio, R. Ducharme and P. Vincent, "A neural probabilistic language model," *Journal of Machine Learning Research*, vol. 3, pp. 1137–1155, February 2003.

[5] V. Siivola and A. Honkela: A state-space method for language modeling, *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2003

13.3 Large vocabulary decoder

The task of the decoder in a speech recognition system is to combine the probabilities and rules given by all the model component to find a word sequence that matches best with the given speech. In order to do this, the decoder should, in principle, consider all possible word sequences, and score them using the acoustic and language models. However, because the number of possible word sequences is extremely large even with small vocabularies, the decoder must concentrate the search effort on the most promising words and prune the improbable sequences in an early stage.

During the past two years, we have been actively developing a large vocabulary decoder [1,2]. Instead of using whole words as recognition units, as traditional speech recognition systems do, our decoder constructs words from smaller units, called morphs. This makes it possible to recognize very large vocabularies with a reasonable number of units, which is important in Finnish, especially. Because natural speech is continuous and does not contain clear word boundaries, the decoder has to consider a possible word boundary after every morph, and use language models to evaluate where the word boundaries are most probable. The decoder puts the combined word sequences in stacks according to their ending times, and only the best sequences are stored for each time instant. In this stack decoding approach, complex language models can be used without hindering the acoustic matching, but the dependence between acoustic models is harder to take into account.

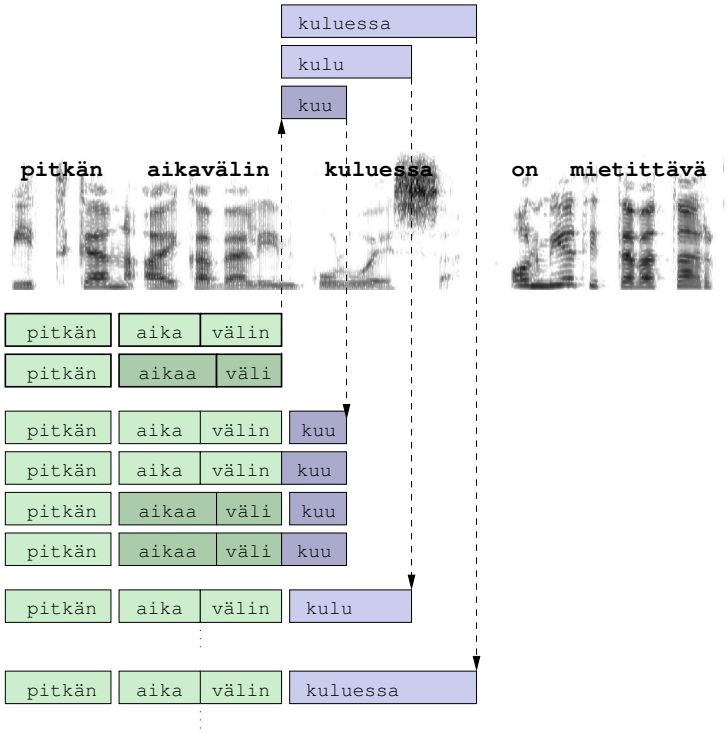


Figure 13.4: The stack decoder expands two hypotheses (bold green boxes) with three acoustically promising most morphs (blue boxes).

References

[1] T. Hirsimäki, "Decoder design for large vocabulary continuous speech recognition system," M.S. thesis, Helsinki University of Technology, Espoo, Finland, 2002.

Chapter 14

SOM in data mining

Esa Alhoniemi, Johan Himberg, Jaakko Hollmén, Sampsa Laine, Pasi Lehtimäki, Kimmo Raivio, Timo Similä, Olli Simula, Miki Sirola, Mika Sulkava, Jarkko Tikka, Juha Vesanto

14.1 Introduction

The Self-Organizing Map (SOM) has proven to be one of the most powerful algorithms in data visualization and exploration. Application areas include various fields of science and technology, e.g., complex industrial processes, telecommunications systems, document and image databases, and even financial applications. The SOM maps the high-dimensional input vectors onto a two-dimensional grid of prototype vectors and orders them. For a human interpreter, the ordered prototype vectors are easier to visualize and explore than the original data. The SOM has been widely implemented in various software tools and libraries, for example, the SOM Toolbox [1].

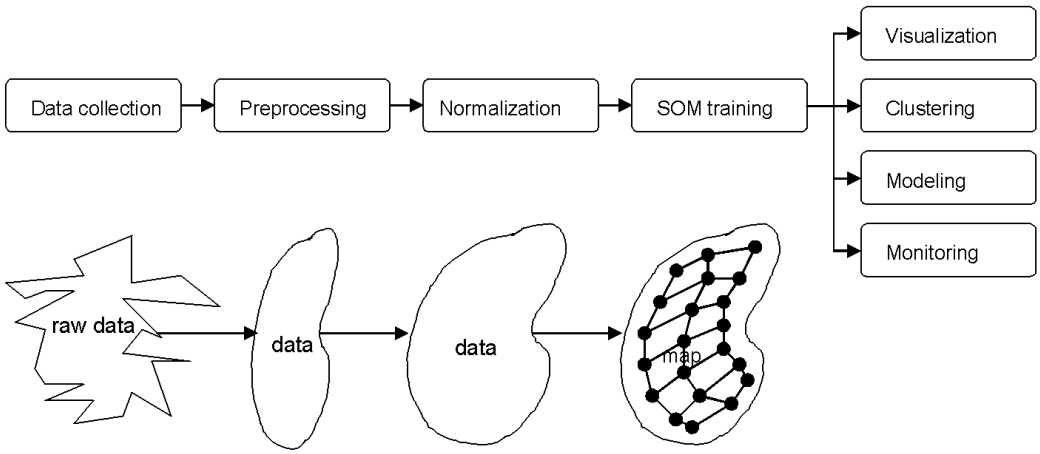


Figure 14.1: Applying the SOM in data mining. Post-processing the SOM extracts qualitative or quantitative information of the data. Visualization and clustering provide qualitative information, while modeling and monitoring give quantitative information resulting in deeper understanding of the system behavior.

The research work has been motivated by a number of practical data mining projects where SOM has been a central data analysis tool [2]. It has become apparent that while the SOM can be used to quickly create a qualitative overview of the data, turning this qualitative information to quantitative characterizations requires a great deal of expertise and manual work. There is no wide consensus or understanding of the methods needed for post-processing of the SOM-based data analysis (see Figure 14.1). The subsequent research has concentrated on devising such methods and on gaining a better understanding of the possibilities, strengths, and weaknesses of the SOM in data exploration.

References

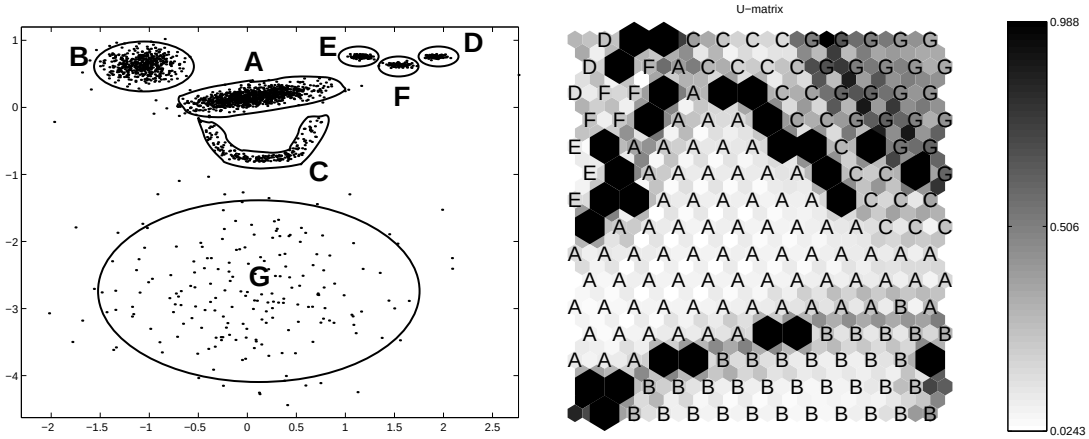
[1] Alhoniemi E., Himberg J. Parhankangas J., Vesanto J., The SOM-toolbox, 2000. Available from `http://www.cis.hut.fi/projects/somtoolbox/`.

[2] Kohonen, T., Self-Organizing Maps, Series in Information Sciences, second edn. 1997, Springer, Heidelberg.

14.2 Clustering of the SOM

Clustering of data is one of the main applications of the Self-Organizing Map (SOM) [1]. U-matrix is one of the most commonly used methods to cluster the SOM visually. However, in order to be really useful, clustering needs to be an automated process. When clusters are identified visually the results may be different when performed by different people. There are several techniques which can be used to cluster the SOM autonomously, but the results they provide do not follow the results of U-matrix very well.

In [2], a clustering approach based on distance matrices was introduced which produces results which are very similar to the U-matrix. It was compared to other SOM-based clustering approaches and found to produce more reliable results. The automated clustering algorithm has also been applied to study the relations of nutrient concentrations in tree needles [3] (see also Section 15.1).



(a) Data with clusters (b) U-matrix and clusters

Figure 14.2: (a) Data set with true clusters indicated with encircled areas and the letters. (b) U-matrix of the data with empty map units shown as black, and clustering result with the letters.

References

[1] Juha Vesanto and Esa Alhoniemi. Clustering of the Self-Organizing Map. *IEEE Transactions on Neural Networks*, 11(2):586–600, March 2000.

[2] Juha Vesanto and Mika Sulkava. Distance matrix based clustering of the self-organizing map. In José R. Dorronsoro, editor, *Artificial Neural Networks - ICANN 2002*, volume 2415 of *Lecture Notes in Computer Science*, pages 951–956, Madrid, Spain, August 2002. Springer.

[3] Mika Sulkava and Jaakko Hollmén. Finding profiles of forest nutrition by clustering of the self-organizing map. In *Proceedings of the Workshop on Self-Organizing Maps (WSOM'03)*, pages 243–248, Hibikino, Kitakyushu, Japan, September 2003.

14.3 Use of operator maps to analyze mobile access network

One of the most flexible extensions of the SOM suitable for the analysis of switching time-series data is the operator map [1]. In operator maps, the map units (or operators) are generalized to be parametric models that are able to describe the interesting behavioral patterns in different parts of the data.

In [2, 3], operator maps were applied in the analysis of downlink traffic performance in a macrocellular network scenario. In the analysis, the relationships between the number of users, the downlink average transmission power and the downlink frame error rate were studied. Several operator maps consisting of 16 local operators in a rectangular lattice of size $[4 \times 4]$ were trained using different types of map operators. In Figure 14.3, an example of an operator map with 16 neuro-fuzzy operators is shown. Each neuro-fuzzy operator provides a linguistic description for the input variable condition in which the quality problems occur, enabling easy human analysis of the dependencies in the data.

1 if nUsr(n) is very high nUsr(n−1) is very high dlTxp(n) is very med dlTxp(n−1) is very med then dlFer is 0.70	5 if nUsr(n) is very med nUsr(n−1) is very high dlTxp(n) is very high dlTxp(n−1) is very med then dlFer is 0.72	9 if nUsr(n) is very high nUsr(n−1) is very high dlTxp(n) is very low dlTxp(n−1) is very med then dlFer is 0.74	13 if nUsr(n) is very high nUsr(n−1) is very high dlTxp(n) is very high dlTxp(n−1) is very med then dlFer is 0.73
2 if nUsr(n) is very med nUsr(n−1) is very high dlTxp(n) is very med dlTxp(n−1) is very med then dlFer is 0.70	6 if nUsr(n) is very low nUsr(n−1) is very high dlTxp(n) is very high dlTxp(n−1) is very med then dlFer is 0.76	10 if nUsr(n) is very high nUsr(n−1) is very^2 high dlTxp(n) is very high dlTxp(n−1) is very med then dlFer is 0.72	14 if nUsr(n) is very med nUsr(n−1) is very high dlTxp(n) is very^2 low dlTxp(n−1) is very high then dlFer is 0.73
3 if nUsr(n) is very high nUsr(n−1) is very^2 high dlTxp(n) is very med dlTxp(n−1) is very med then dlFer is 0.72	7 if nUsr(n) is very low nUsr(n−1) is very High dlTxp(n) is very med dlTxp(n−1) is very High then dlFer is 0.70	11 if nUsr(n) is very low nUsr(n−1) is very high dlTxp(n) is very med dlTxp(n−1) is very High then dlFer is 0.74	15 if nUsr(n) is very high nUsr(n−1) is very med dlTxp(n) is very high dlTxp(n−1) is very low then dlFer is 0.77
4 if nUsr(n) is very med nUsr(n−1) is very med dlTxp(n) is very med dlTxp(n−1) is very^2 high then dlFer is 0.72	8 if nUsr(n) is very med nUsr(n−1) is very high dlTxp(n) is very high dlTxp(n−1) is very med then dlFer is 0.71	12 if nUsr(n) is very high nUsr(n−1) is very low dlTxp(n) is very med dlTxp(n−1) is very med then dlFer is 0.72	16 if nUsr(n) is very med nUsr(n−1) is very high dlTxp(n) is very med dlTxp(n−1) is very med then dlFer is 0.73

Figure 14.3: Rule-based descriptions of the operators.

References

[1] Teuvo Kohonen. Self-Organizing Maps, 3rd edition. Springer, 2001.

[2] Pasi Lehtimäki. Self-Organizing Operator Maps in Complex System Analysis. Master's Thesis, Department of Computer Science and Engireering, Helsinki University of Technology, 2002.

[3] Pasi Lehtimäki, Kimmo Raivio and Olli Simula. Self-Organizing Operator Maps in Complex System Analysis. In *Proceedings of Joint International Conference on Artificial Neural Networks and Neural Information Processing*, pages 622–629, Istanbul, Turkey, June 26 - 29 2003.

14.4 Use of LogSig-scaling to incorporate expert knowledge to SOM-based visualization of GSM-network data

Normalization of data is an important step of data analysis. Since most analysis methods measure distances between data points, a variable having higher variation will dominate the results. A common way to perform normalization is done by subtracting mean and scaling to unit variance each of the variables. Outliers, or equivalently uninteresting parts of the data distribution, reduce weight of interesting parts of the distribution when such normalization is performed. This causes analysis methods to concentrate on wrong issues.

Data collected from operation of GSM-network is studied in order to compare effects of two different normalization methods on information content of SOM trained with normalized data [1]. Process experts of the GSM-network provided value ranges of importance for each of the variables. The proposed normalization method transforms the data by sigmoidal function whose shape is fixed based on auxliary information from the experts. By normalizing the data with the proposed method, the SOM visualizes better overall behavior of the GSM-network, whereas the reference method performing unit variance normalization causes SOM neurons to stretch toward extreme parts of the data distribution. These exreme parts represent severe, but rare problems in network operation. Sammon mappings [2] in Figure 14.4 visualize relative positions of neurons in the two multidimensional SOMs.

References

[1] K. Hätonen, S. Laine, T. Similä. Using the LogSig-function to integrate expert knowledge to Self-Organizing Map (SOM) based analysis. In *Proceedings of the 2003 IEEE International Workshop on Soft Computing in Industrial Applications*, pages 145-150, Binghamton, NY, USA, June 23-25, 2003.

[2] J. Sammon. A nonlinear mapping for data structure analysis. *IEEE Transactions on computers* C-18(5):401-409, 1969.

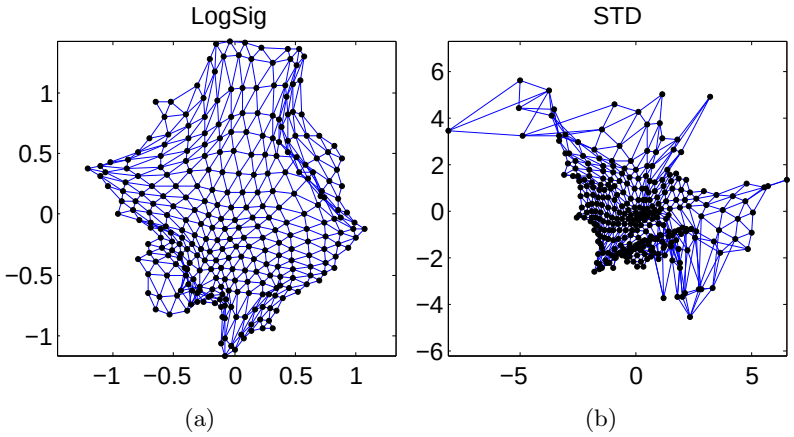


Figure 14.4: Sammon mapping of SOM trained with (a) sigmoidal normalized and (b) unit variance normalized data.

14.5 Analysis of mobile access network

Both 3G and GSM networks produce a huge amount of data. In this project, the Self-Organizing Map has been used to analyse mobile data [1][2]. 3G network data has been generated using a wideband code division multiple access (WCDMA) radio network simulator. The GSM data has 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. Cells of one cluster can be configured using similar parameters.

The method utilizes the SOM algorithm twice when clustering mobile cells. The Self-Organizing Map is used together with some clustering algorithm to cluster data vectors of single mobile cell and to cluster the mobile cell features. 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 so that exact number of clusters can be defined. The input data vectors are classified using labeled SOM codebook vectors.

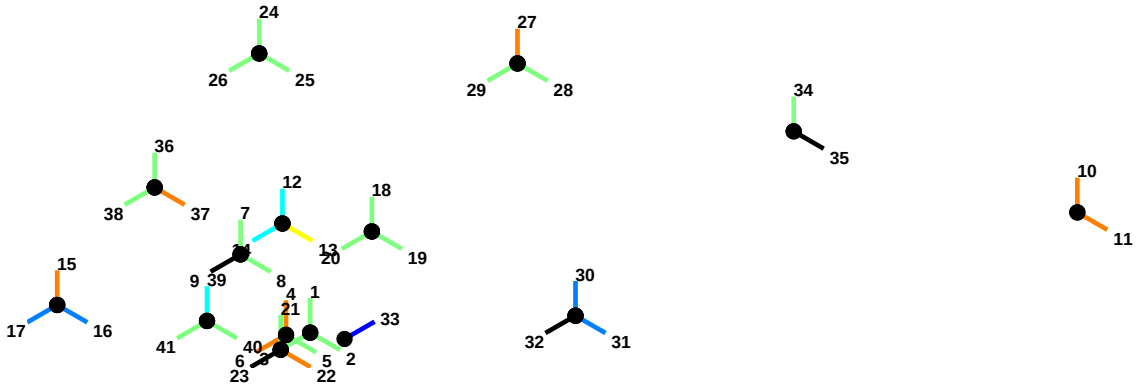


Figure 14.5: Classified GSM cells

For each mobile cell a histogram is computed. 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 a 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 14.5.

In this method, two level clustering procedure has been used because long term cell profiles are desired. At least, in 3G systems this is necessary due to high sampling rates, but also in GSM systems they give us more reliable classification results.

References

[1] Pasi Lehtimäki, Kimmo Raivio, and Olli Simula. Mobile radio access network monitoring using the self-organizing map. In *Proceedings of European Symposium on Artificial Neural Networks*, pages 231–236, Bruges, Belgium, April 24 - 26 2002.

[2] Kimmo Raivio, Olli Simula, Jaana Laiho, and Pasi Lehtimäki. Analysis of mobile radio access network using the self-organizing map. In *Proceedings of the International Symposium on Integrated Network Management*, pages 439–451, Colorado Springs, Colorado, USA, March 24 - 28 2003.

[3] Juha Vesanto and Esa Alhoniemi. Clustering of the self-organizing map. *IEEE Transactions on Neural Networks*, 11(3):586–600, May 2000.

14.6 Impact of R&D on growth quantiles in manufacturing firms

Impact of research and development (R&D) on growth in Finnish manufacturing firms is studied. Growth of a firm is treated as random variable whose distribution is conditional on regressors, namely size of the firm, R&D intensity of the firm, R&D intensity of the industry of the firm and growth of the industry of the firm. Local linear quantile regression model is built in order to capture variation in firm growth given values of the regressors. Since the model is nonparametric, its parameters vary both by quantiles and values of the four regressors. SOM is used in visualization of the quantile regression model. Novelty of this choice is capability to track shapes of the conditional quantiles of firm growth distribution and perform sensitivity analysis for them as function of the regressors. Figure 14.6 shows an illustrative example of SOM in quantile regression visualization.

Results of the study suggest that there exists a relatioship between the conditional quantiles of firm growth and the regressors. Smallest R&D doing firms gain the highest, but also the lowest benefits from firm-level R&D investments independent of other factors. Sensitivity analysis show that only already growing firms gain benefit from increase in firm-level R&D investments in medium and high technology industries, but in low technology industries even non-growing firms may gain benefit. Firms in low technology industries benefit more from increase in industry-level R&D intensity than firms in medium or high technology industries. More detailed results will be published later in a Master's thesis.

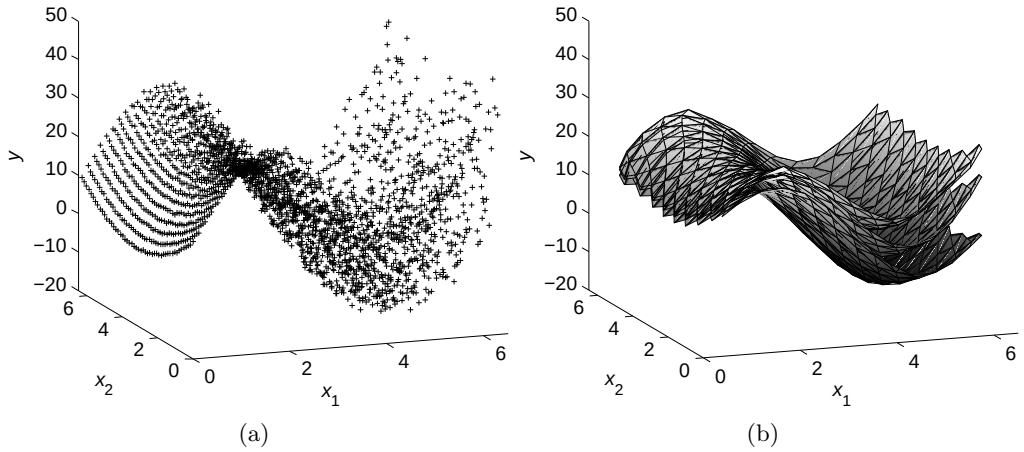


Figure 14.6: (a) Scatter plot of simulated data and (b) three-layered SOM arranged along quantiles 15%, 50% and 85% of random variable $y|\{x_1, x_2\}$.

Chapter 15

Intelligent data engineering

Esa Alhoniemi, Jaakko Hollmen, Johan Himberg, Sampsa Laine, Golan Lampi, Pasi Lehtimäki, Teppo Marin, Jukka Parviainen, Kimmo Raivio, Timo Similä, Olli Simula, Miki Sirola, Mika Sulkava, Jarkko Tikka, Juha Vesanto

15.1 Spatio-temporal analysis of forest nutrition data

Living plants are capable of taking up substances from the environment and using them for the synthesis of their cellular components. Plant nutrients play an integral role in the physiological and biochemical processes of forest ecosystems. Therefore the nutritional status of trees provides an important diagnostic tool for estimating tree condition [1]. In this project, the nutrient concentrations of pine and spruce needles in Finland and Austria between 1987–2000 were studied using different data analysis methods [2]. The aim was to analyze the spatial and temporal distribution of the nutrients and generally find out what kind of internal structure exists in the data. The analysis methods used in [2] were spatial statistics, clustering of the self-organizing map and time series modeling. The work was done in collaboration with the Finnish Forest Research Institute, Parkano Research Station.

The clustering method of the self-organizing map [3] provided new information about the relations of the nutrients between different years and locations. The clustering method was able to represent the structure of the relations of nutrient concentrations in a new informative way. Using the clustering, we were able to divide the measurements into groups [2]. The clustering result of Finland on the geographical map in different years is presented in Figure 15.1. In each group the growth of the needles and the amounts of the nutrients were different and thus, different groups represented different kinds of growing conditions. Using the result of the clustering method, it was possible to construct a temporal model that characterizes the development of the forests of Finland.

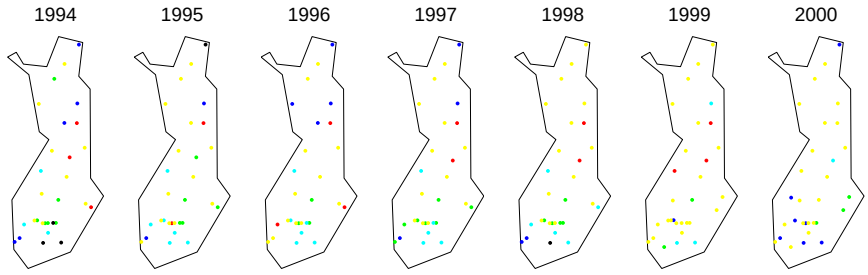


Figure 15.1: Clustering of the measurement stands of Finland for years 1994–2000. Colors indicate different clusters.

References

[1] Sebastiaan Luyssaert, Hannu Raitio, and Alfred Fürst. Forest nutrition at the Finnish and Austrian level I plots in 1987-2000. Technical report, The Finnish Forest Research Institute and Austrian Federal Office Research Centre for Forests, 2003.

[2] Mika Sulkava. Identifying spatial and temporal profiles from forest nutrition data. Master's thesis, Helsinki University of Technology, May 2003.

[3] Juha Vesanto and Mika Sulkava. Distance matrix based clustering of the self-organizing map. In José R. Dorronsoro, editor, *Artificial Neural Networks - ICANN 2002*, volume 2415 of *Lecture Notes in Computer Science*, pages 951–956, Madrid, Spain, August 2002. Springer.

[4] Mika Sulkava and Jaakko Hollmén. Finding profiles of forest nutrition by clustering of the self-organizing map. In *Proceedings of the Workshop on Self-Organizing Maps (WSOM'03)*, pages 243–248, Hibikino, Kitakyushu, Japan, September 2003.

15.2 Using visualization, variable selection and feature extraction to learn from industrial data

Although the engineers of industry have access to process data, they seldom use advanced statistical tools to solve process control problems. Why this reluctance? I believe that the reason is in the history of the development of statistical tools, which were developed in the era of rigorous mathematical modelling, manual computation and small data sets. This created sophisticated tools. The engineers do not understand the principles of these algorithms related. If algorithms are fed with unsuitable data, or are parameterized poorly, they produce biased results, which probably leads an engineer to descregard statistical tools.

My thesis work [1] proposes algorithms that probably do not impress the champions of statistics, but serve process engineers. I advocate the three properties: supervised operation, robustness and understandability. Supervised operation allows and requires the user to explicate the goal of the analysis. Robustness allows the analysis of raw process data. Understandability is essential, as the user must know how to parameterize the algorithm, and how to interpret the results.

To realise a methodology that complies with the above criteria, I studied three types of algorithms: visualization, variable selection and feature extraction. Variable selection helps the user to find relevant variables among the hundreds of variables provided by an automation system; Feature extraction helps the user to mathematically manipulate the variables to surface relevant information; Visualization provides understandable presentation of the results. Figure 15.2 illustrates these three tools together with the three criteria.

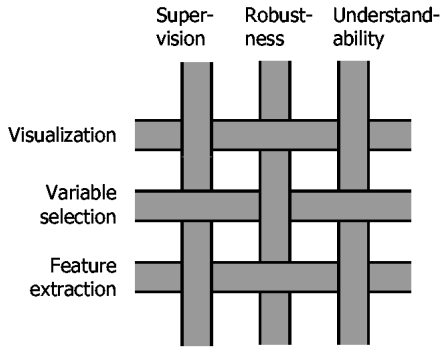


Figure 15.2: The three criteria and tool types discussed in my thesis

I illustrated my approach by analysing an industrial case: the concentrator of the Hitura mine. A significant benefit of algorithmic study of data is efficiency: the manual approach reported in my early publications took approximately six man months to produce; the automated approach of this thesis created comparable results in few weeks.

References

[1] S. Laine, Using visualization, variable selection and feature extraction to learn from industrial data, PhD thesis, Helsinki University of Technology, 2003. Available in `http://lib.hut.fi/Diss/2003/isbn9512266709/`.

15.3 Decision models for computerized decision support

Computerized decision support system prototypes have been developed and summarized in [1], where decision making problem formulation of failure management in safety critical processes was studied. The main application area was the nuclear power plants. Decision support in both the control room and maintenance were covered. One of the models, the use of decision analysis methodology in maintenance problems, is presented in [2].

After these works the objective has been to build decision models for certain applications, mostly for practical purposes, and also try to find out more general decision making principles. This approach leads easily to single case studies that are difficult to generalize. The large amount of possible methodologies and the narrowness of application areas are also known difficulties.

To find out general principles from separate case studies, to formulate more comprehensive decision concepts, and to build more general decision models are difficult tasks. While such studies produce tested models and concepts, evaluation of these results is difficult, because there are no competent measures for such purposes. The only really clear result is the decision support achieved in each particular case.

How to utilize data analysis in computerized decision support systems has been outlined. A prototype is being built to demonstrate how to utilize Self-Organizing Map in a computerized decision support system.

An old decision case that has been analyzed with rule-based methodologies in [1] has been solved with multi-criteria decision analysis method in [3]. A Comparison with the elder case has been made in the analysis. The problem is to choose the right control action in a situation where a leak has appeared in the primary circuit of a BWR nuclear power plant.

Decision concepts have been reviewed and a conceptual decision model has been built by case-based means [4]. This model utilizes rule-based methodologies, numerical algorithms and procedures, statistical methodologies including distributions, and visual support. Probability models are used in handling uncertainties.

References

[1] Sirola M. Computerized decision support systems in failure and maintenance management of safety critical processes. VTT Publications 397. Espoo, Finland, 1999.

[2] Laakso K., Sirola M., Holmberg J. Decision modelling for maintenance and safety. International Journal of Condition Monitoring and Diagnostic Engineering Management. Birmingham, England, July 1999.

[3] Sirola, M. Applying decision analysis method in process control problem in accident management situation. International Conference on Systems Science. Wroclaw, Poland, September 2001.

[4] Sirola M. Using conceptual decision model in a case study. International Conference on Knowledge-Based Intelligent Information & Engineering Systems (KES'2003). Oxford, United Kingdom, September 2003.

15.4 Context awareness

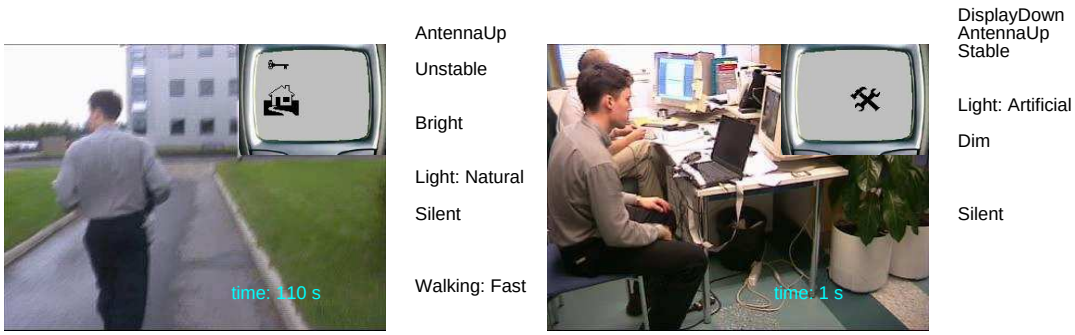
Context awareness [4] has become a major topic in human-computer interaction. The need for context awareness is especially large in mobile communications, where the communication situations can vary a lot. A mobile terminal is often expected to enable connections all the time. At the same time, it should not irritate the user by signalling in a wrong way at the wrong moment, or by requiring constant attention to keep it working in the right way for the situation. In addition, the hand-held terminals are becoming more and more sophisticated in their function yet smaller in their size. The interaction could be made easier and less intruding if the mobile device recognized the user's current context and adapted its functions accordingly.

Information of user's preferences is obtained from the logs of different applications, e.g., calling, messaging, using calendar, or profiling. Piece of ambient information can be obtained by directly monitoring the user's physical environment using on-board sensors and information of user's location. The operating network itself can offer information, e.g., on location. Setting explicit information sources, context tags, located in a short range network is another approach.

The device can infer parts of the context of the user from features extracted from on-board measurements of acceleration, noise level, luminosity, humidity, etc. In [1],[2], we have consider context recognition by fusing and clustering these context features using a recently introduced method, the Symbol Clustering Map (SCM) [3]. As such, it can be used for finding static patterns but a suitable transformation of the data allows identifying also temporal patterns. The recognized clusters/segments can then serve as "higher-level contexts" that show which combinations of the basic features form common patterns in the data.

Fig. 15.3 presents an user, the context features and the recognized context in two different situations. The context is presented here as a user interface profile. In this case, common contexts are recognized unsupervised by the SCM from training data. However, the labeling (deciding the profile) is done afterwards by the user, and the selection of the profiles/actions is based on a lookup table. A future aim is that the terminal would also learn to suggest applications according to user's spontaneous actions in different situations.

Publications [1, 2] are joint work with Dr. John Adrian Flanagan in Nokia Research Center, Helsinki, Finland and Dr. Jani Mäntyjärvi in VTT Technical Research Centre of Finland, Oulu, Finland.



(a) (b)

Figure 15.3: In panel (a) SCM has recognized a "walking outdoors" context based on the active features listed to the right of the image. "Keypad lock on" and "Outdoors profile" have been activated according the the action lookup table. In Panel (b) a "working profile" is launched due to the office context.

References

[1] J.A. Flanagan, J. Himberg, and J. Mäntyjärvi. A Hierarchical Approach to Learning Context and Facilitating User Interaction in Mobile Devices. In *Proceedings of Artificial Intelligence in Mobile System 2003 (AIMS 2003).* (in conjunction with Ubicomp 2003, October 12, Seattle, USA.)

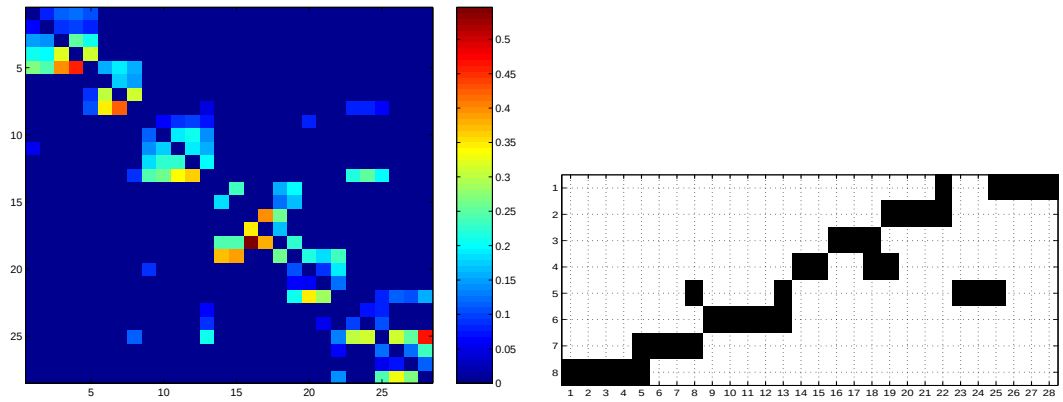
[2] J. Himberg, J. A. Flanagan, and J. Mäntyjärvi. Towards Context Awareness Using Symbol Clustering Map. In *Proceedings of the Workshop on Self-Organizing Maps (WSOM'03)*, pp. 249–254, Hibikino, Kitakyushu, Japan, September 2003.

[3] A. Flanagan. Unsupervised Cluster Discovery using the Self-Organizing Map. In *Proc. of International Conference on Artificial Neural Networks and Neural Information Processing (ICANN/ICONIP 2003)*, pp. 9-12, 2003, Istanbul, Turkey.

[4] A.K. Dey and G.D. Abowd and D. Salber. A Conceptual Framework and a Toolkit for Supporting the Rapid Prototyping of Context-Aware Applications. *Human-Computer Interaction*, 16(2–4):97–166, 2001.

15.5 Dependency trees from industrial time-series data

Industrial processes generate large masses of multivariate, noisy time-series data. In exploring the database, the analyst is interested in looking at the structure of the data set, or more specifically dependencies among the variables. Our problem is to seek for dependencies between the N variables in the data set. The dependencies are defined through multiple linear regression models, which are estimated from the data. Before model fitting, time-series are denoised using the Wavelet transform. In model fitting, one variable at the time is the dependent variable and rest of the variables are possible regressors. Sparse regression algorithms are used to select the best regressors among all candidates and estimate the corresponding regression coefficients. Bootstrap is also applied on the selection and the estimation. The relative weight of the each regressor is computed from the bootstrap replications of the regression coefficients. The relative weight of the regressor is a measure of belief that the regressor belongs to the estimated linear model. These relative weights are thresholded to yield graphs, some graph operations are performed to define dependent variables. Taken together, the method defines a dependency tree, or possibly a dependency forest. More detailed results will be reported in a Master's thesis.



(a) (b)

Figure 15.4: Examples from an artificial data set. In (a), the relative energies of the sparse linear model coefficients estimated from the bootstrapped data sets are shown. In (b), dependencies between variables are defined using thresholding and operations on the resulting graph.

Chapter 16

Other projects

16.1 PRIMA—Proactive information retrieval by adaptive models of users' attention and interests

Samuel Kaski, Jarkko Salojärvi, Eerika Savia, Kai Puolamäki

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 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.

In information retrieval 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 the 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). It started in 2003, and the first results are on modeling of eye movements.

Predicting relevance from eye movements

We measure eye movements during reading, and based on this implicit feedback, try to infer how relevant the document is to the user. Eye movements have earlier been used as alternative input devices in human-computer interfaces (e.g. [5]), and recently in a proactive dictionary which becomes automatically activated [1]. To our knowledge, they have not been used in information retrieval before.

The main challenges are that (i) the signal is complex and very noisy, and (ii) interestingness or relevance is higly subjective and thus hard to define. We started the project by feasibility studies to find out whether the problems are solvable.

We constructed a controlled experimental setting in which it is known which documents are relevant, and then tried to learn relevance from measured eye movement patterns. The user was instructed to find an answer to a specific question, and then shown a set of document titles (Fig. 16.1), of which some were known to be relevant.

In the first feasibility study [3] we extracted a set of standard features [2] from eye movements for each word and combined them to title-specific feature vectors. The two goals of analysing the data were to find out whether relevance can be estimated in this simplified setup using standard features, and which features were important in predicting the relevance. The data was explored with unsupervised methods (Principal Component Analysis and Self-Organizing Maps), and their supervised versions, Linear Discriminant Analysis (LDA) and SOM that learns metrics (cf. Section on Learning metrics).

The results were encouraging; even a simple linear classifier was able to determine relevance clearly better than by chance (80.5% vs. 63%), and a subset of five features was sufficient. There were also many non-linear effects in the data, implicating that a better discrimination is to be expected with a non-linear classifier.

Classification accuracy is also likely to improve when the temporal structure of the data is taken into account. We have started work on Hidden Markov Models (HMMs), which

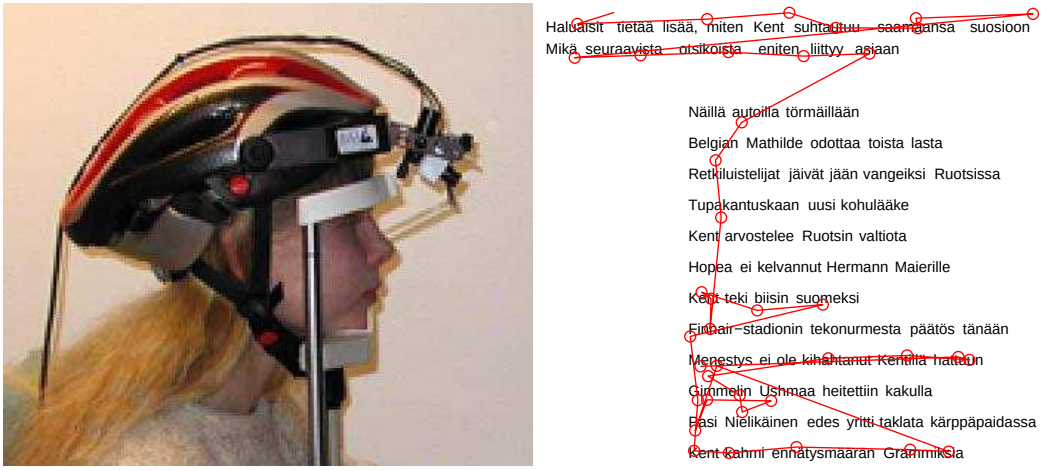


Figure 16.1: The experimental setup. Left: The eye movements of the user are being tracked with a head-mounted eye tracker. The tracker consists of a helmet with two cameras; one monitors the eye and the other one the visual field of the subject. Right: The eye movement pattern during reading plotted 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.

have earlier been used for segmenting the low-level eye movement signal to detect focus of attention (see [6]) and for implementing (fixed) models of cognitive processing [4]. First results of applying HMMs to our problem setting show improvement of the classification accuracy from 69.2% (using LDA) to 75.8% (NIPS Machine Learning Meets the User Interface workshop, December 2003).

References

[1] Aulikki Hyrskykari, Päivi Majaranta, and Kari-Jouko Räihä. Proactive response to eye movements. In *Proc. INTERACT'03*. 2003. To appear.

[2] Keith Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3):372–422, 1998.

[3] Jarkko Salojärvi, Ilpo Kojo, Jaana Simola, and Samuel Kaski. Can relevance be inferred from eye movements in information retrieval ? In *Proceedings of the Workshop on Self-Organizing Maps (WSOM'03)*, pages 261–266, Hibikino, Kitakyushu, Japan, September 2003.

[4] Dario D. Salvucci and John R. Anderson. Automated eye-movement protocol analysis. *Human-Computer Interaction*, 16:39–86, 2001.

[5] David J. Ward and David J.C. MacKay. Fast hands-free writing by gaze direction. *Nature*, 418:838, 2002.

[6] Chen Yu and Dana H. Ballard. A multimodal learning interface for grounding spoken language in sensory perceptions. In *Proc. ICMI'03*. ACM, 2003. To appear.

16.2 Data analysis using a tree-shaped neural network

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.

Classical computer science has a long history of dealing with data sets. One of the most common approaches is the *divide and conquer* approach, where a large problem is separated into smaller subproblems. Another way to approach the problem is using different kinds of *search trees*, which efficiently index the data.

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. The first realization of this research is *The Evolving Tree* [1].

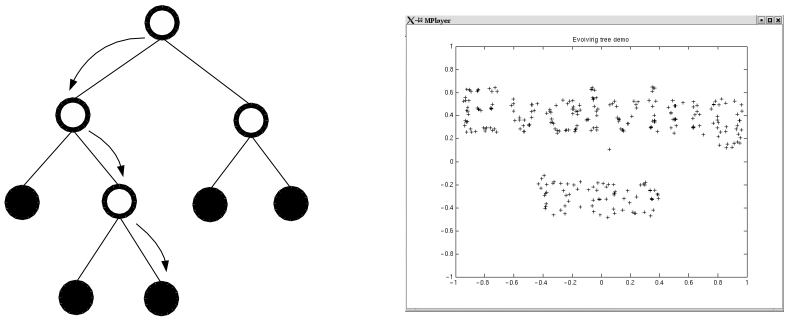


Figure 16.2: The general architecture of the Evolving Tree and an example of adaptation to data.

Figure 16.2 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 16.2 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 artificial and real world data indicate that the Evolving Tree could be applied to several problems, such as pattern recognition, data mining, density estimation, and exploratory data analysis.

References

[1] J. Pakkanen The Evolving Tree, a new kind of self-organizing neural network, in *Proceedings of the workshop on Self-Organizing Maps '03*, pages 311–316, September 2003, Kitakyushu, Japan.

16.3 Computational model of visual attention

Teuvo Kohonen

By means of simple modeling approaches, an explicit explanation has been given in this work to the following phenomena: 1. Automatic *activation* of a subset of visual signal paths, equivalent to an "attentional window," such that the width of the window is defined by the relative variances of the visual signals. 2. *Narrowing* of the "attentional window" when small saccadic eye movements, voluntary or involuntary, are made. This effect can be shown to ensue from the same model, when the primary signals are further high-pass filtered. 3. *Shifting* of the "attentional window" when strong or novel stimuli (distractors) occur eccentrically in the visual field.

The channel organization

Consider the circular subareas in Fig. 16.3, which delineates a simplified model retina. In this kind of mapping the small foveal areas and the large peripheral areas are thought to project into areas of equal size in the higher parts of the brain. Then we may imagine that the signal paths starting at the retina and ending up on the visual cortex are organized in spatially ordered, functionally separate *channels* corresponding to the small circles in Fig. 16.3. A channel is here identified with a set of signal paths, the *transmittance* of which is controlled by a common *control circuit*. The transmittances of the channels are assumed to have soft shoulders, e.g., Gaussian.

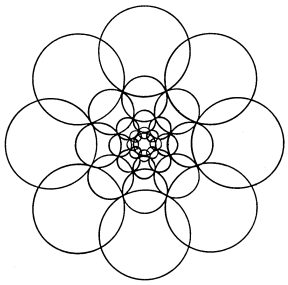


Figure 16.3: Placement of the channels over a hypothetical model retina, around the fovea. A control circuit with corresponding (effective) diameter is associated with each circle.

Assume now that the control circuit of each channel is able to analyze some kind of information content in its incoming signals. The control circuits shall also be able to compare their information contents and mutually *compete* on the permission for activation.

Consider that if we want to compare the information content of subareas relating to such an inhomogeneous sampling system as the retina, any information measure should be related to the resolution of vision in the corresponding subarea.

If the cross section of each channel is then partitioned into an equal number of subfields, if the intensity of the picture is averaged over each such subfield, and if the *variance* of these averaged values is then taken, we obtain a robust measure that is independent of the width of the channel and describes variations of the signal intensity at the given resolution. Let us call this kind of "information measure" the *resolution-related variance*. Notwithstanding, since the absolute variations are slightly different in the light and dark areas of the image, it has further turned out, for photographic images at least, to be most effective to divide the variance by the average of the signal values in the channel.

The sampling grid

In simulations, photographic images were used where the pixels were defined in an orthogonal grid. The resolution-related variance in each channel shown in Fig. 16.3 was computed by placing a *sampling grid* over the channel (Fig. 16.4); the diameter of the sampling grid shall be selected to correspond to the diameter of the due channel, and thus around the assumed direction of the gaze the sampling grid shall be smaller and have fewer pixels, whereas the diameter of the grid shall be selected wider and more pixels must be covered with increasing distance from the direction of the gaze. A constant number, e.g., seven subsets of pixels over each sampling grid were defined, and the averages $av_i, i = 1 \ldots 7$ of the pixels over these subsets were computed. The resolution-related variance for each sampling grid was evaluated by computing the variance of the av_i. After that, the variance was divided by the average of all pixels of this grid. The figure so obtained defines the variable *Variance* in Eq. (16.1).

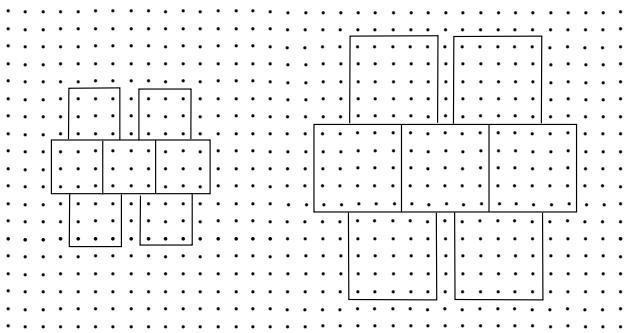


Figure 16.4: Two examples of the sampling grids used for the control of gating of signal transfer in simulations. The small dots correspond to pixels. Over each of the seven square areas, the average $av_i, i = 1 \ldots 7$ of the pixels is computed, whereafter the variance of the av_i is evaluated, and the variance is further divided by the average of pixels over the seven squares.

Optimal width of a channel

Before discussing the system of channels as delineated in Fig. 16.3, it may be interesting to find out how an *optimal width* of a channel, concentrated at a particular location of the image, is determined by the resolution-related variance.

Consider that we try channels of varying width at a certain location of the image. We are looking for the width of the control grid that maximizes the normalized variance of the local averages av_i of the pixel values, denoted *Variance*. Let us call the image data vector **Image**. Let **Grid**(w) mean the choice for the grid with width w; then the "optimal" width w_o is defined to be

$$w_o = \arg\max_w \{ \mathit{Variance}[\mathbf{Grid}(w), \mathbf{Image}] \} . \tag{16.1}$$

A robust optimization of w_o in Eq. (16.1) was carried out over a discrete set of five sampling grids, with their widths varying from 10 to 80 pixels, respectively.

In the first series of simulations illustrated in Fig. 16.5 we demonstrate the "optimal" width of the attentional window, when the gaze was directed at various objects of different widths; the fish, the palm, and the telephone pole, respectively.

Figure 16.5: Demonstration of the opening of attentional windows, the widths of which were automatically determined by the structures present in the area around the gaze. First and third picture: original images. The rest of the pictures show attentional windows, when the gaze was directed to one fish, the palm, and the telephone pole, respectively.

Narrowing of the attentional window

The next phenomenon that is explainable by the optimization approach is the *narrowing of the attentional window* when the gaze is *moved*, voluntarily or involuntarily, by a small amount.

Let us assume that every sampling grid, to some extent, has also high-pass filter properties, i.e., it enhances transient (phasic) values of the signals it samples. Let these temporal variations of the signals ensue from the shifts of the gaze, i.e., translations of the input image over the sampling grids.

Consider the spatial frequencies of the images: if the translation is small, the absolute value of the difference is approximately proportional to the Euclidean norm of its gradient, in which high spatial frequencies are enhanced in proportion to the frequency. In the evaluation of the optimal width w_o from Eq. (16.1), the variances computed from the av_i for the difference image thus decrease with the width of the grid, too, and the optimal width w_o is decreased.

When only a fraction of the previous image is subtracted from the new image, a similar shift of w_o towards smaller values, although a weaker one, can be seen. This effect is then reflected as narrowing of the attentional window. In Fig. 16.6 a sequence of images is

shown, where the subtracted fraction was 50 per cent of each previous image.

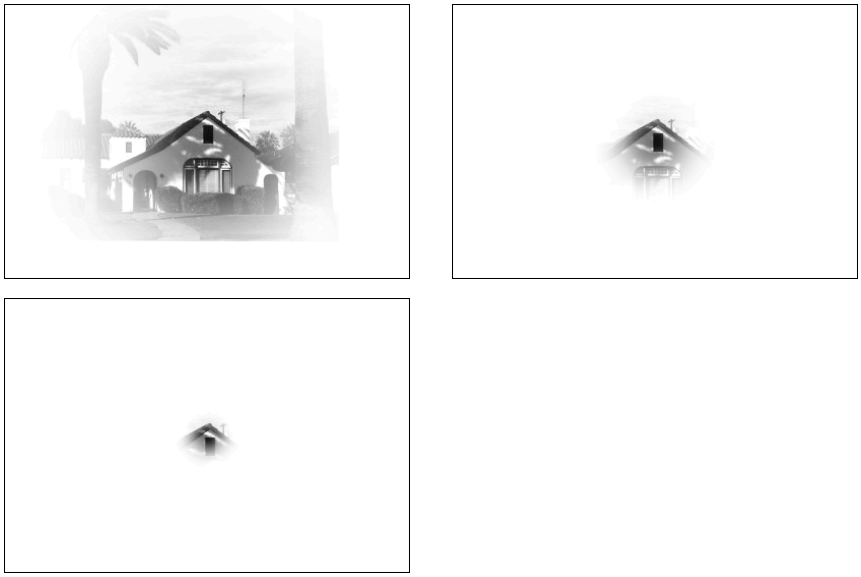


Figure 16.6: Automatic narrowing of the attentional window, when the variances were computed on the basis of images from which 50 per cent of the previously sampled translated image was subtracted. The three pictures form a sequence, in which the gaze was shifted in steps, the size of which became successively smaller.

Attentional window as an activated subset of channels

Finally we shall consider the more complete "biological" case in which the set of channels is fixed and their sizes and positions were defined in Fig. 16.3. For each channel, a sampling grid of corresponding diameter is associated.

Instead of looking for the optimized width of the channel as before, we thus now keep the positions and widths of the channels *fixed* and try to determine a *combination* of k activated channels over which the normalized resolution-related variance of the av_i is highest. In this way, while most of the channels are located eccentrically with respect to the direction of the gaze, the combination of the activated channels defines a more or less symmetric (usually noncircular) attentional window.

In the simulation presented in Fig. 16.7 we thus use the 33-channel "retina" of Fig. 16.3 and let four highest-variance channels define the attentional window. As can be seen, the four channels together tend to emphasize a part of the visual field where some meaningful pattern is present.

It is also discernible that if the variance in the central part of the visual field is low, prominent eccentric patterns tend to *attenuate* weaker parts of the visual field, which can then be interpreted as the *distraction* of attention by the prominent eccentric objects.

References

[1] T. Kohonen, "Modeling of automatic capture and focusing of visual attention," *PNAS*, vol. 99, pp. 9813-9818, 2002.

[2] T. Kohonen, "A computational model of visual attention," Proc. IJCNN'03 (CD-ROM).

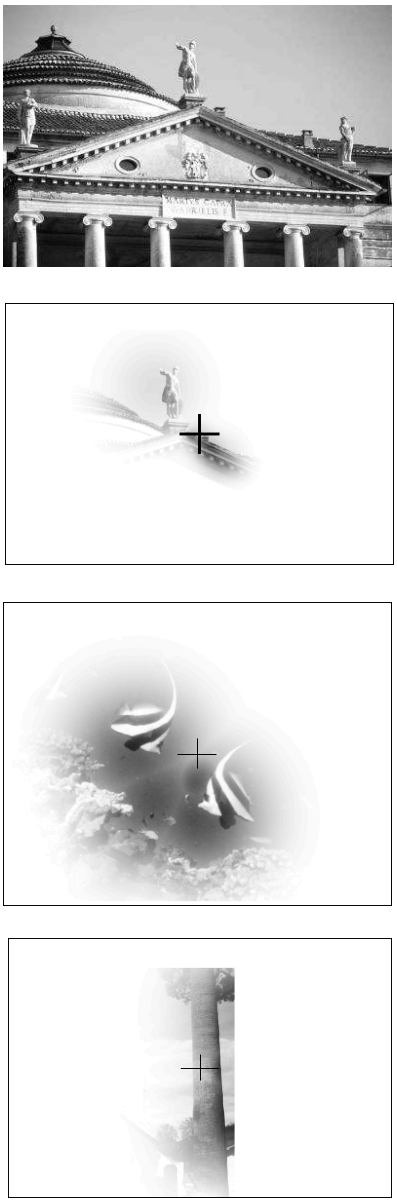


Figure 16.7: Examples of attentional windows spanned by a combination of four activated channels. The black cross indicates the direction of the gaze. The first picture is another original image, of which a part (the statue) is selected and emphasized in the second picture. In the third picture (cf. the first picture in Fig. 16.5, a butterfly-formed attentional window, compassing two of the fishes, is opened. In the lowest picture (cf. the third picture in Fig. 16.5), the form of the resulting attentional window is oblong and the window stretches along the trunk of the tree.

Publications of the Neural Networks Research Centre

Publications are in alphabetical order by the first author.

2002

1. Aksela, M., Girdziusas, R., Laaksonen, J., Oja, E. & Kangas, J. Class-Confidence Critic Combining. *Proc. of 8th International Workshop on Frontiers in Handwriting Recognition (IWFHR-8)*, Niagara-on-the-Lake, Ontario, Canada, August 6-8, 2002. pp. 201-206.

2. Ala-Siuru, P. & Kaski, S. (eds.) *STeP 2002 - Intelligence, the Art of Natural and Artificial. Proceedings of the 10th Finnish Artificial Intelligence Conference*, Oulu, Finland, Sept. 15-17, 2002. Helsinki 2002, Finnish Artificial Intelligence Society.

3. Amari, S.-I., Hyvärinen, A., Lee, S.-Y., Lee, T.-W. & Sanchez A., V.D. Blind Signal Separation and Independent Component Analysis. *Neurocomputing*, 2002. Vol. 49, pp. 1-5.

4. Bingham, E., Kuusisto, J. & Lagus, K. ICA and SOM in Text Document Analysis. *Proc. of 25th ACM SIGIR 2002 International Conference on Research and Development in Information Retrieval*, Tampere, Finland, August 11-15, 2002. pp. 361-362.

5. Bingham, E., Mannila, H. & Seppänen, J. K. Topics in 0-1 Data. *Proc. of 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Edmonton, Alberta, Canada, July 23-26, 2002. pp. 450-455.

6. Brandt, S., Laaksonen, J. & Oja, E. Statistical Shape Features for Content-Based Image Retrieval. *Journal of Mathematical Imaging and Vision*, 2002. Vol. 17, No. 2, pp. 187-198.

7. Creutz, M. & Lagus, K. Unsupervised Discovery of Morphemes. *Proc. of Workshop on Morphological and Phonological Learning of ACL-02*, Philadelphia, Pennsylvania, USA, July 11, 2002. pp. 21-30.

8. Díaz, I. & Hollmén, J Residuals Generation and Visualization for Understanding Novel Process Conditions. *Proc. of IEEE 2002 International Joint Conference on Neural Networks (IJCNN'02)*, Honolulu, Hawaii, USA, May 12-17, 2002. pp. 2070-2075.

9. Federley, M., Alhoniemi, E., Laitila, M., Suojärvi, M. & Ritala, R. State Management for Process Monitoring, Diagnostics and Optimization. *Pulp & Paper Canada*, 2002. Vol. 103, No. 2, pp. 40-43.

10. Girolami, M. Latent Variable Models for the Topographic Organisation of Discrete and Strictly Positive Data. *Neurocomputing*, 2002. Vol. 48, No. 1-4, pp. 185-198.

11. Girolami, M. Mercer Kernel-Based Clustering in Feature Space. *IEEE Trans. on Neural Networks*, 2002. Vol. 13, No. 3, pp. 780-784 .

12. Girolami, M. Orthogonal Series Density Estimation and the Kernel Eigenvalue Problem. *Neural Computation*, 2002. Vol. 14, pp. 669-688.

13. Honkela, A. Speeding Up Cyclic Update Schemes by Pattern Searches. *Proc. of 9th International Conference on Neural Information Processing (ICONIP'02)*, Singapore, November 18-22, 2002. pp. 512-516.

14. Hoyer, P. O. Non-Negative Sparse Coding. *Proc. of IEEE Workshop on Neural Networks for Signal Processing*, Martigny, Switzerland, September 4-6, 2002. pp. 557–565.

15. Hoyer, P. O. & Hyvärinen, A. A Multi-Layer Sparse Coding Network Learns Contour Coding from Natural Images. *Vision Research*, 2002. Vol. 42, No. 12, pp. 1593-1605.

16. Hoyer, P. O. & Hyvärinen, A. Sparse Coding of Natural Contours. *Neurocomputing*, 2002. Vol. 44–46, pp. 459-466.

17. Hurri, J. & Hyvärinen, A. A Novel Temporal Generative Model of Natural Video as an Internal Model in Early Vision. *Proc. of First Int. Workshop on Generative-Model-Based Vision*, Copenhagen, Denmark, June 2, 2002. pp. 37–41.

18. Hurri, J. & Hyvärinen, A. Receptive Fields Similar to Simple Cells Maximize Temporal Coherence in Natural Video. *Proc. of Int. Conf. on Artificial Neural Networks - ICANN 2002*, Madrid, Spain, August 28-30, 2002. pp. 33-38.

19. Hyvärinen, A. Activity Bubbles and Natural Image Sequences. *Proc. of Int. Conf. on Knowledge-Based Intelligent Information and Engineering Systems (KES2002)*, Crema, Italy, Sept. 16-18, 2002. pp. 397-401.

20. Hyvärinen, A. An Alternative Approach to Infomax and Independent Component Analysis. *Neurocomputing*, 2002. Vol. 44-46, pp. 1089-1097.

21. Hyvärinen, A. Topography as a Property of the Natural Sensory World. *Natural Computing*, 2002. Vol. 1, No. 2–3, pp. 185-198.

22. Hyvärinen, A. & Inki, M. Estimating Overcomplete Independent Component Bases from Image Windows. *Journal of Mathematical Imaging and Vision*, 2002. Vol. 17, pp. 139–152.

23. Hyvärinen, A. & Raju, K. Imposing Sparsity on the Mixing Matrix in Independent Component Analysis. *Neurocomputing*, 2002. Vol. 49, pp. 151-162.

24. Iivarinen, J., Kaski, S. & Oja, E. (Eds.) *Neljännesvuosisata Hatutusta: Hahmontunnistustutkimus Suomessa 1977-2002.* Espoo, Finland 2002, Suomen hahmontunnistustutkimuksen seura ry.

25. Iivarinen, J. & Pakkanen, J. Content-Based Retrieval of Defect Images. *Proc. of Advanced Concepts for Intelligent Vision Systems*, Ghent, Belgium, September 9-11, 2002. pp. 62-67.

26. Iivarinen, J., Pakkanen, J. & Rauhamaa, J. Content-Based Image Retrieval in Surface Inspection. *Proc. of 7th International Conference on Control, Automation, Robotics and Vision*, Singapore, December 3-6, 2002. pp. 182-188.

27. Ilvesmäki, A. & Iivarinen, J. Defect Image Classification with MPEG-7 Descriptors. *Proc. of STeP2002*, Oulu, Finland, December 16-17, 2002. pp. 196-202.

28. Inki, M. & Hyvärinen, A. Two Approaches to Estimation of Overcomplete Independent Component Bases. *Proc. of Int. Joint Conference on Neural Networks (IJCNN 2002)*, Honolulu, Hawaii, USA, May 12-17, 2002. pp. 454-459.

29. Jokinen, K., Kerminen, A., Kaipainen, M., Jauhiainen, T., Turunen, M., Hakulinen, J., Kuusisto, J. & Lagus, K. Adaptive Dialogue Systems - Interaction with Interact. *Proc. of 3rd SIGdial Workshop on Discourse and Dialogue of ACL-02*, Philadelphia, NJ, July 11-12, 2002. pp. 64-73.

30. Kaban, A. & Girolami, M.A. Fast Extraction of Semantic Features from a Latent Semantic Indexed Text Corpus. *Neural Processing Letters*, 2002. Vol. 15, pp. 31-43.

31. Kaski, S. Learning Metrics. *Proc. of 3rd Conference of the International Society for Ecological Informatics*, Grottaferrata, Italy, August 26-30, 2002. p. 42.

32. Kaski, S., Sinkkonen, J. & Klami, A. Regularized Discriminative Clustering. Espoo, Finland: Helsinki University of Technology, 2002. 8 p. (Publications in Computer and Information Science Report A67).

33. Kersting, K., Raiko, T. & De Raedt, L. Logical Hidden Markov Models (Extended Abstract). *First European Workshop on Graphical Models (PGM-02)*, 6-8 November 2002, Cuenca, Spain. pp. 99-107.

34. Kersting, K., Raiko, T., Kramer, S. & De Raedt, L. Towards Discovering Structural Signatures of Protein Folds based on Logical Hidden Markov Models (Extended abstract). *Work-in-Progress Reports of the Twelfth International Conference on Inductive Logic Programming (ILP-2002)*, Sydney, Australia, July 9-11, 2002.

35. Klami, A., Peltonen, J. & Kaski, S. Accurate Self-Organizing Maps in Learning Metrics. *Proc. of STeP -2002, Suomen Tekoälytutkimuksen Päivät, Finnish AI Conference*, Oulu, Finland, December 15-17, 2002. pp. 41-49.

36. Kohonen, T. Modeling of Automatic Capture and Focusing of Visual Attention. *PNAS*, 2002. Vol. 99, No. 15, pp. 9813-9818.
http://www.pnas.org/cgi/doi/10.1073/pnas.152318799

37. Kohonen, T. & Somervuo, P. How to Make Large Self-Organizing Maps for Nonvectorial Data. *Neural Networks*, 2002. Vol. 15, No. 8-9, pp. 945-952.

38. Koskela, M., Laaksonen, J. & Oja, E. Implementing Relevance Feedback as Convolutions of Local Neighborhoods on Self- Organizing Maps. *Proc. of International Conference on Artificial Neural Networks*, Madrid, Spain, August 2002. pp. 981-986.

39. Koskela, M., Laaksonen, J. & Oja, E. MPEG-7 Descriptors in Content-Based Image Retrieval with PicSOM System. *Proc. of 5th International Conference on Visual Information System*, HsinChu, Taiwan, March 11-13, 2002. pp. 247-258.

40. Koskela, M., Laaksonen, J. & Oja, E. Using MPEG-7 Descriptors in Image Retrieval with Self-Organizing Maps. *Proc. of International Conference on Pattern Recognition*, Quebec, Canada, August 2002. vol. 2, pp. 1049-1052.

41. Kurimo, M. Puheentunnistus. In: Iivarinen, J., Kaski, S. & Oja, E. (Eds.), *Neljännesvuosissata Hatutusta - Hahmontunnistustutkimus Suomessa 1977 - 2002.* Espoo, Finland 2002, Suomen hahmontunnistustutkimuksen seura ry - Pattern Recognition Society of Finland, pp. 115-123.

42. Kurimo, M. Thematic Indexing of Spoken Documents by Using Self-Organizing Maps. *Speech Communication*, 2002. Vol. 38, No. 1-2, pp. 29-44.

43. Kurimo, M. & Lagus, K. An Efficiently Focusing Large Vocabulary Language Model. *Proc. of International Conference on Artificial Neural Networks (ICANN'02)*, Madrid, Spain, August 28-30, 2002. pp. 1068-1073.

44. Könönen, V. Focused Crawling Using Fictitious Play. *Proc. of Third International Conference on Intelligent Data Engineering and Automated Learning (IDEAL'02)*, Manchester, UK, August 12-14, 2002. pp. 186-192.

45. Laaksonen, J., Koskela, M. & Oja, E. PicSOM - Self-Organizing Image Retrieval with MPEG-7 Content Descriptions. *IEEE Trans. on Neural Networks*, 2002. Vol. 13, No. 4, pp. 841-853.

46. Lagus, K. Text Retrieval Using Self-Organized Document Maps. *Neural Processing Letters*, 2002. Vol. 15, No. 1, pp. 21-29.

47. Lagus, K., Airola, A. & Creutz, M. Data Analysis of Conceptual Similarities of Finnish Verbs. *Proc. of 24th Annual Meeting of the Cognitive Science Society (CogSci 2002)*, Fairfax, Virginia, USA, August 7-10, 2002. pp. 566-571.

48. Lagus, K. & Kurimo, M. Language Model Adaptation in Speech Recognition using Document Maps. *Proc. of IEEE Workshop on Neural Networks for Signal Processing (NNSP'02)*, Martigny, Switzerland, September 4-6, 2002. pp. 627-636.

49. Lagus, K. & Kuusisto, J. Topic Identification in Natural Language Dialogues Using Neural Networks. *Proc. of 3rd SIGdial Workshop on Discourse and Dialogue of ACL-02*, Philadelphia, NJ, USA, July 11-12, 2002. pp. 95-102.

50. Laiho, J., Raivio, K., Lehtimäki, P., Hätönen, K. & Simula, O. Advanced Analysis Methods for 3G Cellular Networks. Espoo, Finland: Helsinki University of Technology, 2002. (Publications in Computer and Information Science Report A65).

51. Laine, S. Finding the Variables of Interest. *Minerals Engineering*, 2002. Vol. 15, pp. 167-176.

52. Laine, S. Selecting the Variables that Train a Self-Organizing Map (SOM) which Best Separates Predefined Clusters. *Proc. of 9th International Conference on Neural Information Processing (ICONIP'02)*, Singapore, November 18-22, 2002.

53. Lehtimäki, P., Raivio, K. & Simula, O. Mobile Radio Access Network Monitoring Using the Self-Organizing Map. *Proc. of European Symposium on Artificial Neural Networks*, Bruges, Belgium, April 23-25, 2002. pp. 231–236.

54. Lindblom, N., Heiskala, H., Kaski, M., Leinonen, L. & Laakso, M.-L. Sleep Fragmentation in Mentally Retarded People Decreases with Increasing Daylength in Spring. *Chronobiology International*, 2002. Vol. 19, No. 2, pp. 441-459.

55. Lindén, K. & Lagus, K. Word Sense Disambiguation in Document Space. *Proc. of 2002 IEEE Int. Conference on Systems, Man and Cybernetics*, Tunisia, October 6-9, 2002. (CD-ROM)

56. Nikkilä, J., Törönen, P., Kaski, S., Venna, J., Castrén, E. & Wong, G. Analysis and Visualization of Gene Expression Data Using Self-Organizing Maps. *Neural Networks*, 2002. Vol. 15, pp. 953-966.

57. Oja, E. Convergence of the Symmetrical FastICA Algorithm. *Proc. of 9th Int. Conf. on Neural Information Processing (ICONIP'02)*, Singapore, Nov. 18-22, 2002.

58. Oja, E. Finding Hidden Factors Using Independent Component Analysis. *Proc. of 13th European Conf. on Machine Learning*, Helsinki, Finland, Aug. 19-23, 2002. p. 505.

59. Oja, E. Independent Component Analysis. *Proc. of 2nd Int. Conf. on Hybrid Intelligent Systems*, Santiago, Chile, Dec. 1-4, 2002.

60. Oja, E. Keynote Talk: Independent Component Analysis: Recent Advances. *Proc. of 6th Int. Conf. on Cognitive and Neural Systems*, Boston, Massach., USA, May 30 - June 1, 2002. p. 7.

61. Oja, E. Unsupervised Learning in Neural Computation. *Theoretical Computer Science*, 2002. Vol. 287, No. 1, pp. 187-207.

62. Oja, M., Nikkilä, J., Törönen, P., Castrén, E. & Kaski, S. Learning Metrics for Visualizing Gene Functional Similarities. *Proc. of STeP2002, Finnish AI Conference, Suomen Tekoälytutkimuksen Päivät*, Oulu, Finland, December 15-17, 2002. pp. 31-40.

63. Oja, M., Nikkilä, J., Törönen, P., Wong, G., Castrén, E. & Kaski, S. Exploratory Clustering of Gene Expression Profiles of Mutated Yeast Strains. In: Zhang, W. and Shmulevich, I. (eds.), *Computational and Statistical Approaches to Genomics*. Boston 2002, Kluwer, pp. 65-78.

64. Ollikainen, V., Bäckström, C. & Kaski, S. Electronic Editor: Automatic Content-based Sequential Compilation of Newspaper Articles. *Neurocomputing*, 2002. Vol. 43, pp. 91-106.

65. Pakkanen, J. & Iivarinen, J. Evaluating SOM as an Index in Content-Based Image Retrieval. *Proc. of STeP2002*, Oulu, Finland, December 16-17, 2002. pp. 182-188.

66. Peltonen, J., Klami, A. & Kaski, S. Learning More Accurate Metrics for Self-Organizing Maps. *Proc. of International Conference on Artificial Neural Networks, ICANN 2002*, Madrid, Spain, August 27- 31, 2002. Springer, pp. 999-1004.

67. Peltonen, J., Sinkkonen, J. & Kaski, S. Discriminative Clustering of Text Documents. *Proc. of 9th International Conference on Neural Information Processing (ICONIP'02)*, Singapore, November 18-22, 2002.

68. Raiko, T., Kersting, K., Karhunen, J. & De Raedt, L. Bayesian Learning of Logical Hidden Markov Models. *Proc. of Finnish Artificial Intelligence Conference (STeP2002)*, December 16-17, 2002, Oulu, Finland. pp. 64- 71.

69. Raju, K. & Ristaniemi, T. ICA-RAKE-Switch for Jammer Cancellation in DS-CDMA Array Systems. *Proc. of 2002 IEEE Int. Symp. on Spread Spectrum Techniques and Applications (ISSTA 2002)*, Prague, Czech Republic, September 2-5, 2002. pp. 638-642.

70. Raju, K., Ristaniemi, T., Karhunen, J. & Oja, E. Suppression of Bit-Pulsed Jammer Signals in a DS-CDMA Array Systems Using Independent Component Analysis. *Proc. of IEEE Int. Symp. on Circuits and Systems (ISCAS2002)*, Phoenix, Arizona, USA, May 26-29, 2002. pp. I-189–192.

71. Ristaniemi, T., Raju, K. & Karhunen, J. Jammer Mitigation in DS-CDMA Array Systems Using Independent Component Analysis. *Proc. of IEEE Int. Conf. on Communications (ICC2002)*, New York City, NY, USA, April 28 - May 2, 2002.

72. Ristaniemi, T., Raju, K., Karhunen, J. & Oja, E. Jammer Cancellation in DS-CDMA Array Systems: Pre and Post Switching of ICA and RAKE. *Proc. of 2002 IEEE Int. Symposium on Neural Networks for Signal Processing (NNSP)*, Martigny, Switzerland, September 4-6, 2002. pp. 495-504.

73. Saastamoinen, K., Könönen, V. & Luukka, P. A Classifier Based on the Fuzzy Similarity in the Lukasiewicz Structure with Different Metrics. *Proc. of 2002 World Congress on Computational Intelligence (WCCI'02)*, Honolulu, Hawaii, USA, May 12-17, 2002.

74. Salojärvi, J. & Kaski, S. Mixture Density from Autonomous Experts. *International Journal of Knowledge-Based Intelligent Engineering Systems*, 2002. Vol. 6, pp. 48-55.

75. Seppänen, J. K., Hollmén, J., Bingham, E. & Mannila, H. Nonnegative Matrix Factorization on Gene Expression Data. *Bioinformatics 2002*, Bergen, Norway, April 4-7, 2002. poster 49. http://www.ii.uib.no/bio2002/abstractsbio2002.pdf

76. Siivola, V., Hirsimäki, T. & Kurimo, M. Äännemallien vertailua jatkuvassa suuren sanaston puheentunnistuksessa. *22nd Fonetiikan päivät*, Espoo, Finland, August 29-31, 2002. 2002, Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing,

77. Sinkkonen, J. & Kaski, S. Clustering Based on Conditional Distributions in an Auxiliary Space. *Neural Computation*, 2002. Vol. 14, pp. 217-239.

78. Sinkkonen, J., Kaski, S. & Nikkilä, J. Discriminative Clustering: Optimal Contingency Tables by Learning Metrics. *Proc. of 13th European Conference on Machine Learning (ECML'02)*, Helsinki, Finland, August 19-23, 2002. Springer, pp. 418-430.

79. Sirola, M. Applying Decision Analysis Method in a Special Accident Management Process Control Problem. *Proc. of IASTED International Conference on Intelligent Systems and Control*, Tsukuba, Japan, October 2-4, 2002. 6 p.

80. Somervuo, P. Speech Modeling Using Variational Bayesian Mixture of Gaussians. *Proc. of 7th International Conference on Spoken Language Processing (ICSLP)*, Denver, Colorado USA, Sept. 16-20, 2002. pp. 1245-1248.

81. Somervuo, P. Two-Level Phoneme Recognition Based on Successive Use of Monophone and Diphone Models. *Proc. of EUSIPCO, XI European Signal Processing Conference*, Sept. 3-6, 2002, Toulouse, France.

82. Valpola, H., Honkela, A. & Karhunen, J. An Ensemble Learning Approach to Nonlinear Dynamic Blind Source Separation Using State-Space Models. *Proc. of Int. Joint Conf. on Neural Networks (IJCNN2002)*, Honolulu, Hawaii, USA, May 12-17, 2002. pp. 460-465.

83. Valpola, H. & Karhunen, J. An Unsupervised Ensemble Learning Method for Nonlinear Dynamic State-Space Models. *Neural Computation*, 2002. Vol. 14, No. 11, pp. 2647-2692.

84. Valpola, H., Raiko, T. & Karhunen, J. Constructing Graphical Models for Bayesian Ensemble Learning from Simple Building Blocks. *Learning Workshop*, Snowbird, Utah, USA, April 2-5, 2002. 2 p.

85. Valpola, H. & Särelä, J. A Fast Semi-Blind Source Separation Algorithm. Espoo, Finland: Helsinki University of Technology, 2002. 4 p. (Publications in Computer and Information Science Report A66).

86. Vesanto, J. & Hollmén, J. An Automated Report Generation Tool for the Data Understanding Phase. *Proc. of First International Conference on Hybrid Intelligent Systems (HIS'01)*, Adelaide, Australia, Dec. 11-12, 2001. Heidelberg 2002, Physica-Verlag, pp. 611-625.

87. Vesanto, J. & Hollmén, J. An Automated Report Generation Tool for the Data Understanding Phase. In: Abraham, A., Jain. L. C. & Kacprzyk, J. (eds.), *Recent Advances in Intelligent Paradigms and Applications, Studies in Fuzziness and Soft Computing (Vol. 113)*. Heidelberg 2002, Physica (Springer) Verlag,

88. Vesanto, J. & Sulkava, M. Distance Matrix Based Clustering of the Self-Organizing Map. *Proc. of International Conference on Artificial Neural Networks - ICANN 2002*, Madrid, Spain, August 28- 30, 2002. pp. 951-956.

89. Viitaniemi, V. & Laaksonen, J. Browsing an Electronic Mail-Order Catalogue with PicSOM. *Proc. of STeP -2002, SuomenTekoälytutkimuksen Päivät*, Oulu, Finland, December 15-17, 2002. pp. 170- 181.

90. Vuori, V. Clustering Writing Styles with a Self-Organizing Map. *Proc. of 8th International Workshop on Frontiers in Handwriting Recognition (IWFHR-8)*, Niagara-on-the-Lake, Ontario, Canada, August 6-8, 2002. pp. 345-350.

91. Vuori, V. & Laaksonen, J. A Comparison of Techniques for Automatic Clustering of Handwritten Characters. *Proc. of 16th International Conference on Pattern Recognition*, Quebec, Canada, August 2002. pp. 168-171.

92. Vuori, V., Laaksonen, J. & Kangas, J. Influence of Erroneous Learning Samples on Adaptation in On-Line Handwriting Recognition. *Pattern Recognition*, 2002. Vol. 35, pp. 915-925.

2003

1. Aksela, M. Comparison of Classifier Selection Methods for Improving Committee Performance. *Proc. of 4th International Workshop on Multiple Classifier Systems, MCS2003*, Guildford, Surrey, United Kingdom, June 11-13, 2003. pp. 84-93.

2. Aksela, M., Girdziusas, R., Laaksonen, J., Oja, E. & Kangas, J. Methods for Adaptive Combination of Classifiers with Application to Recognition of Handwritten Characters. *Int. J. of Document Analysis and Recognition*, 2003. Vol. 6, No. 1, pp. 23-41.

3. Alila-Johansson, A., Eriksson, L., Soveri, T. & Laakso, M.-L. Serum Cortisol Levels in Goats Exhibit Seasonal but not Daily Rhythmicity. *Chronobiology International*, 2003. Vol. 20, pp. 65-79.

4. Bingham, E., Kaban, A. & Girolami, M. Topic Identification in Dynamical Text by Complexity Pursuit. *Neural Processing Letters*, 2003. Vol. 17, No. 1, pp. 69-83.

5. Creutz, M. Unsupervised Segmentation of Words Using Prior Distributions of Morph Length and Frequency. *Proc. of 41st Annual Meeting of the Association of Computational Linguistics, ACL-03*, Sapporo, Japan, July 7-12, 2003. pp. 280-287.

6. Flanagan, J.A., Himberg, J. & Mäntyjärvi, J. A Hierarchical Approach to Learning Context and Facilitating User Interaction in Mobile Devices. *Proc. of Mobile System 2003 (AIMS 2003)*, (in conjunction with Ubicomp 2003), Seattle, USA, October 12, 2003.

7. Funaro, M., Oja, E. & Valpola, H. Independent Component Analysis for Artefact Separation in Astrophysical Images. *Neural Networks*, 2003. Vol. 16, No. 3-4, pp. 469-478.

8. Gross, A., Joffe, G., Joutsiniemi, S., Nyberg, P., Rimon, R. & Appelberg, B. Decreased production of reactive oxygen species by blood monocytes caused by clopazine correlates with EEG slowing schizophrenic patients. *Neuropsychobiology*, 2003. Vol 46, No. 2, pp. 73-77.

9. Hacioglu, K., Pellom, B., Ciloglu, T., Ozturk, O., Kurimo, M. & Creutz, M. On Lexicon Creation for Turkish LVCSR. *Proc. of 8th European Conference on Speech Communication and Technology (Eurospeech)*, Geneva, Switzerland, September 1-4, 2003. pp. 1165-1168.

10. Hacioglu, K., Pellom, B., Ciloglu, T., Ozturk, O., Kurimo, M. & Creutz, M. Word Splitting for Turkish LVCSR. *Proc. of Turkish Signal Processing Conference (SIU 2003)*, Istanbul, Turkey, June 18-20, 2003.

11. Himberg, J., Flanagan, J. A. & Mäntyjärvi, J. Towards Context Awareness Using Symbol Clustering Map. *Proc. of Workshop on Self-Organizing Maps (WSOM2003)*, Kitakyushu, Japan, September 11-14, 2003. pp. 249-254.

12. Himberg, J. & Hyvärinen, A. Icasso: Software for Investigating the Reliability of ICA Estimates by Clustering and Visualization. *Proc. of 2003 IEEE Workshop on Neural Networks for Signal Processing (NNSP2003)*, Toulouse, France, September 17-19, 2003. pp. 259-268.

13. Honkela, A. & Valpola, H. On-line Variational Bayesian Learning. *Proc. of Fourth International Symposium on Independent Component Analysis and Blind Signal Separation (ICA 2003)*, Nara, Japan, April 2003. pp. 803-808.

14. Honkela, A., Valpola, H. & Karhunen, J. Accelerating Cyclic Update Algorithms for Parameter Estimation by Pattern Searches. *Neural Processing Letters*, 2003. Vol. 17, No. 2, pp. 191-203.

15. Honkela, T. Kielen sävyt ja tulkinnan moniulotteisuus. *Synteesi*, 2003, No. 2, pp. 58-62.

16. Honkela, T., Honkela, J., Myyryläinen, M. & Tuulos, V. Adaptive and Visual Map Interface for Document Collections. *Adjunct Proceedings of Human-Computer Interaction International*, Crete, Greece, June 22-27, 2003. pp. 119-120.

17. Honkela, T., Hynnä, K.I. & Knuuttila, T. Framework for Modeling Partial Conceptual Autonomy of Adaptive and Communicating Agents. *Proc. of CogSci2003, 25th Annual Meeting of the Cognitive Science Society*, Boston, MA, USA, July 31 - Aug. 2, 2003.

18. Honkela, T., Kurimo, M., Lagus, K., Lantz, V. & Oja, E. Unsupervised Learning in Human-Computer Interfaces. *Adjunct Proceedings of Human-Computer Interaction International*, Crete, Greece, June 22-27, 2003. pp. 121-122.

19. Honkela, T. & Winter, J. Simulating Language Learning in Community of Agents Using Self-Organizing Maps. Espoo, Finland: Helsinki University of Technology, 2003. (Publications in Computer and Information Science Report A71).

20. Honkela, T., Hyvärinen, A. & Väyrynen, J. Emergence of Linguistic Representations by Independent Component Analysis. Espoo, Finland: Helsinki University of Technology, 2003. (Publications in Computer and Information Science Report A72).

21. Hoyer, P. Modeling Receptive Felds with Non-Negative Sparse Coding. *Neurocomputing*, 2003. Vol. 52-54, pp. 547-552.

22. Hoyer, P. & Hyvärinen, A. Interpreting Neural Response Variability as Monte Carlo Sampling of the Posterior. In: *Advances in Neural Information Processing Systems 15* (NIPS 2002, Vancouver, Canada, December 10-12, 2002). Cambridge, MA 2003, MIT Press, pp. 277-284.

23. Hurri, J. & Hyvärinen, A. A Two-Layer Temporal Generative Model of Natural Video Exhibits Complex-Cell-Like Pooling of Simple Cell Outputs. In: *Computational Neuroscience: Trends in Research 2003* (11th Annual Computational Neuroscience Meeting, Chicago, Illinois, USA, July 2002). Amsterdam 2003, Elsevier, pp. 553-559.

24. Hurri, J. & Hyvärinen, A. Simple-Cell-Like Receptive Fields Maximize Temporal Coherence in Natural Video. *Neural Computation*, 2003. Vol. 15, No. 3, pp. 663-691.

25. Hurri, J. & Hyvärinen, A. Temporal and Spatiotemporal Coherence in Simple-Cell Responses: A Generative Model of Natural Image Sequences. *Network: Computation in Neural Systems*, 2003. Vol. 14, No. 3, pp. 527-551.

26. Hurri, J. & Hyvärinen, A. Temporal Coherence, Natural Image Sequences, and the Visual Cortex. In: *Advances in Neural Information Processing Systems 15* (NIPS 2002, Vancouver, Canada, December 10-12, 2002). Cambridge, MA 2003, MIT Press, pp. 141-148.

27. Hyvärinen, A. & Bingham, E. Connection Between Multilayer Perceptrons and Regression Using Independent Component Analysis. *Neurocomputing*, 2003. Vol. 50, No. C, pp. 211-222.

28. Hyvärinen, A., Hoyer, P. O. & Hurri, J. Extensions of ICA as Models of Natural Images and Visual Processing. *Proc. of International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003)*, Nara, Japan, April 1-4, 2003. pp. 963-974.

29. Hyvärinen, A., Hurri, J. & Väyrynen, J. Bubbles: A Unifying Framework for Low-Level Statistical Properties of Natural Image Sequences. *Journal of the Optical Society of America A*, 2003. Vol. 20, No. 7, pp. 1237-1252.

30. Hyvärinen, A. & Kano, Y. Independent Component Analysis for Non-Normal Factor Analysis. In: *New Developments in Psychometrics*, 2003 (Proc. International Meeting of the Psychometric Society (IMPS2001), Osaka, Japan, July 15-19, 2001). pp. 649-656.

31. Hätönen, K., Laine, S. & Similä, T. Using the LogSig-function to Integrate Expert Knowledge to Self-Organizing Map (SOM) Based Analysis. *Proc. of IEEE International Workshop on Soft Computing in Industrial Applications (SMCia/03)*, Binghamton, NY, USA, June 23-25, 2003. pp. 145-150.

32. Iivarinen, J., Ilvesmäki, A. & Pakkanen, J. Evaluation of Features for Defect Image Retrieval. *Proc. of 3rd IASTED International Conference on Visualization, Imaging, and Image Processing*, Benalmádena, Spain, September 8-10, 2003. vol. II, pp. 680-685.

33. Ilin, A. & Valpola, H. On the Effect of the Form of the Posterior Approximation in Variational Learning of ICA Models. *Proc. of 4rd International Conference on Independent Component Analysis and Blind Signal Separation, ICA 2003*, Nara, Japan, April 2003. pp. 915-920.

34. Inki, M. Examining the Dependencies Between ICA Features of Image Data. *Supplementary Proc. ICANN/ICONIP 2003*, Istanbul, Turkey, June 2003. pp. 298-301.

35. Inki, M. ICA Features of Image Data in One, Two and Three Dimensions. *Proc. of Fourth International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003)*, Nara, Japan, April 1-4, 2003. pp. 861-866.

36. Jutten, C. & Karhunen, J. Advances in Nonlinear Blind Source Separation. *Proc. of 4th Int. Symp. on Independent Component Analysis and Blind Signal Separation (ICA2003)*, Nara, Japan, April 2003. pp. 245-256.

37. Jutten, C., Karhunen, J., Almeida, L. & Harmeling, S. Technical Report on Separation Methods for Nonlinear Mixtures. Grenoble: Institut National Polytechnique de Grenoble, 2003. 23 p. + 14 attachm. (Deliverable D29 of the European Joint Project BLISS (Blind Source Separation and Applications, IST-1999-14190)). http://www.lis.inpg.fr/pages_perso/bliss/

38. Karhunen, J., Meinecke, F., Valpola, H. & Ziehe, A. Final Technical Report on BSS Models and Methods for Non-Independent BSS. Grenoble: Institut National Polytechnique de Grenoble, 2003. 23 p. + 4 attachm. (Deliverable D21 of the European Joint Project BLISS (Blind Source Separation and Applications, IST-1999-14190)). http://www.lis.inpg.fr/pages_perso/bliss/

39. Kaski, S. Discriminative Clustering. *Proc. of 54th Session of the International Statistical Institute*, Berlin, Germany, August 13-20, 2003. International Statistical Institute, pp. 270-273.

40. Kaski, S., Nikkilä, J. & Kohonen, T. Methods for Exploratory Cluster Analysis. In: Szczepaniak, P. S., Segovia, J., Kacprzyk, J. & Zadeh, L. A. (eds.), *Intelligent Exploration of the Web*. Berlin 2003, Springer, pp. 136-151.

41. Kaski, S., Nikkilä, J., Oja, M., Venna, J., Törrönen, P. & Castrén, E. Trustworthiness and Metrics in Visualizing Similarity of Gene Expression. *BMC Bioinformatics*, 2003. Vol. 4, No. 48.

42. Kaski, S. & Peltonen, J. Informative Discriminant Analysis. *Proc. of ICML-2003, the Twentieth International Conference on Machine Learning*, Washington, D.C., USA, August 21-24, 2003. Menlo Park, CA 2003, AAAI Press, pp. 329-336.

43. Kaski, S. & Sinkkonen, J. Discriminative Clustering: Vector Quantization in Learning Metrics. *Proc. of 26th Annual Conference of the Gesellschaft für Klassifikation*, Mannheim, Germany, July 22-24, 2002. Berlin 2003, Springer, pp. 456-463.

44. Kaski, S., Sinkkonen, J. & Klami, A. Regularized Discriminative Clustering. *Proc. of 2003 IEEE International Workshop on Neural Networks for Signal Processing*, Toulouse, France, September 17-19, 2003. New York, NY 2003, IEEE, pp. 289-298.

45. Kaynak, O., Alpaydin, E., Oja, E. & Xu, L. (Eds.) *Artificial Neural Networks and Neural Information Processing - ICANN/ICONIP 2003*, Lecture Notes in Computer Science 2714. Berlin, Germany 2003, Springer.

46. Kersting, K., Raiko, T. & De Raedt, L. A Structural GEM for Learning Logical Hidden Markov Models. *Working Notes of the Second KDD-Workshop on Multi-Relational Data Mining (MRDM-03)*, Washington DC, USA, August 27, 2003.

47. Kersting, K., Raiko, T., Kramer, S. & De Raedt, L. Towards Discovering Structural Signatures of Protein Folds based on Logical Hidden Markov Models. *Proc. of Pacific Symposium on Biocomputing, PSB-2003*, Kauai, Hawaii, January 3-7, 2003.

48. Kiviniemi, V., Kantola, J.-H., Jauhiainen, J., Hyvärinen, A. & Tervonen, O. Independent Component Analysis of Nondeterministic fMRI Signal Sources. *NeuroImage*, 2003. Vol. 19, No. 2, pp. 253-260.

49. Kohonen, T. Self-Organized Maps of Sensory Events. *Philosophical Transactions of Royal Society of London A*, 2003. Vol. 361, pp. 1177-1186.

50. Kohonen, T. A Computational Model of Visual Attention. *Proc. of Int. Joint Conf. on Neural Networks, IJCNN'03*, Portland, Oregon, USA, July 20-24, 2003 (CD-ROM).

51. Kortejärvi, H., Malkki, J. & Pajunen, P. Novel Application of Bayesian Approach to Level A in Vitro-in Vivo Correlation (IVIVC) Model for Levosimendan Modified-Release Capsules. *Proc. of 2003 AAPS Annual Meeting and Exposition*, Salt Lake City, Utah, USA, October 26-30, 2003. American Association of Pharmaceutical Scientists.

52. Koskela, M. & Laaksonen, J. Using Long-Term Learning to Improve Efficiency of Content-Based Image Retrieval. *Proc. of 3rd International Workshop on Pattern Recognition in Information Systems (PRIS 2003)*, Angers, France, April 2003. pp. 72-79.

53. Koskela, M., Laaksonen, J. & Oja, E. Inter-Query Relevance Learning in PicSOM for Content-Based Image Retrieval. *Supplementary Proc. of International Conference on Artificial Neural Networks (ICANN'03)*, Istanbul, Turkey, June 26-28, 2003. pp. 520-523.

54. Könönen, V. Asymmetric Multiagent Reinforcement Learning. *Proc. of 2003 WIC International Conference on Intelligent Agent Technology (IAT-2003)*, Halifax, Canada, October 13-17, 2003. IEEE Press, pp. 336-342.

55. Könönen, V. Gradient Based Method for Symmetric and Asymmetric Multiagent Reinforcement Learning. *Proc. of Fourth International Conference on Intelligent Data Engineering and Automated Learning (IDEAL'03)*, Hong Kong, China, March 21-23, 2003. 2003, Springer-Verlag, pp. 68-75.

56. Könönen, V. Policy Gradient Method for Multiagent Reinforcement Learning. *Proc. of Second International Conference on Computational Intelligence, Robotics and Autonomous Systems (CIRAS 2003)*, Singapore, December 15-18, 2003. (CD-ROM)

57. Laaksonen, J., Koskela, M. & Oja, E. Probability Interpretation of Distributions on SOM Surfaces. *Proc. of Workshop on Self-Organizing Maps (WSOM'03)*, Hibikino, Japan, September 11-14, 2003. pp. 77-82.

58. Laine, S. Automatic Extraction of Simple Features from Process Data. *Proc. of International Joint Conference on Artificial Neural Networks and Neural Information Processing (ICANN/ICONIP 2003)*, Istanbul, Turkey, June 26-29, 2003. pp. 134-137.

59. Lee, T.-W., Cardoso, J.-F., Oja, E. & Amari, S. Introduction to Special Issue on Independent Component Analysis. *J. of Machine Learning Res.*, 2003. Vol. 3, No. 8, pp. 1175-1176.

60. Lehtimäki, P., Raivio, K. & Simula, O. Self-Organizing Operator Maps in Complex System Analysis. *Proc. of International Joint Conference on Artificial Neural Networks and Neural Information Processing (ICANN/ICONIP 2003)*, Istanbul, Turkey, June 26-29, 2003. pp. 622-629.

61. Leinonen, L., Joutsiniemi, S.-L., Laakso, M.-L., Lindblom, N. & Kaski, M. Automatic Blink Detection: A Method for Differentiation of Wake and Sleep of Intellectually Disabled and Healthy Subjects in Long-Term Ambulatory Monitoring. *Sleep*, 2003. Vol. 26, No. 4, pp. 473-479.

62. Leinonen, L., Laakso, M.-L., Carlson, S. & Linnankoski, I. Shared Means and Meaning in Vocal Expression of Man and Macaque. *Logoped Phoniatr Vocol*, 2003. Vol. 28, pp. 53-61.

63. Lindén, K. Word Sense Disambiguation with THESSOM. *Proc. of Workshop on Self-Organizing Maps, WSOM'03 - Intelligent Systems and Innovational Computing*, Kitakyushu, Japan, September 11-14, 2003. (CD-ROM)

64. Nikkilä, J., Roos, C., Sinkkonen, J. & Kaski, S. Associative Clustering to Find Dependencies Between Expression Profiles and Transcription Factor Binding. *Proc. of European Conference on Computational Biology, ECCB2003*, Paris, France, September 27-30, 2003. pp. 433-434.

65. Oja, E. & Plumbley, M. Blind Separation of Positive Sources Using Non-Negative PCA. *Proc. of 4th Int. Symp. on Independent Component Analysis and Blind Source Separation*, Nara, Japan, April 1-4, 2003. pp. 11-16.

66. Oja, M., Kaski, S. & Kohonen, T. Bibliography of Self-Organizing Map (SOM) Papers: 1998-2001 Addendum. *Neural Computing Surveys*, 2003. Vol. 3, pp. 1-156. http://www.soe.ucsc.edu/NCS/

67. Oja, M., Somervuo, P., Kaski, S. & Kohonen, T. Clustering of Human Endogenous Retrovirus Sequences with Median Self-Organizing Map. *Proc. of Workshop on Self-Organizing Maps (WSOM'03)*, Hibikino, Japan, September 11-14, 2003. pp. 134-139. (CD-ROM)

68. Pakkanen, J. The Evolving Tree, a New Kind of Self-Organizing Neural Network. *Proc. of Workshop on Self-Organizing Maps (WSOM'03)*, Hibikino, Kitakyushu, Japan, September 11-14, 2003. pp. 311-316.

69. Pakkanen, J. & Iivarinen, J. Content-Based Retrieval of Surface Defect Images with MPEG-7 Descriptors. *Sixth International Conference on Quality Control by Artificial Vision, Proc. SPIE 5132*, 2003. pp. 201-208.

70. Pakkanen, J., Ilvesmäki, A. & Iivarinen, J. Defect Image Classification and Retrieval with MPEG-7 Descriptors. *Proc. of 13th Scandinavian Conference on Image Analysis*, Göteborg, Sweden, June 29-July 2, 2003. pp. 349-355.

71. Parviainen, J. & Simula, O. Misinterpretations When Using Matlab in DSP Education. *Proc. of 2003 Finnish Signal Processing Symposium, Finsig '03*, Tampere, Finland, May 19, 2003. pp. 273-276.

72. Peltonen, J., Klami, A. & Kaski, S. Learning Metrics for Information Visualization. *Proc. of Workshop on Self-Organizing Maps (WSOM'03)*, Hibikino, Kitakyushu, Japan, September 2003. pp. 213-218.

73. Peltonen, J., Sinkkonen, J. & Kaski, S. Finite Sequential Information Bottleneck (fsIB). Espoo, Finland: Helsinki University of Technology, 2003. (Publications in Computer and Information Science Report A74).

74. Pham, D.-T., Ziehe, A., Karhunen, J. & Jutten, C., Final Technical Report on Linear ICA. Grenoble: Institut National Polytechnique de Grenoble, 2003. 19 p. + 5 attachm. (Deliverable D19 of the European Joint Project BLISS (Blind Source Separation and Applications, IST-1999-14190)). http://www.lis.inpg.fr/pages_perso/bliss/

75. Plumbley, M. & Oja, E. A "Non-Negative PCA" Algorithm for Independent Component Analysis. *IEEE Trans. on Neural Networks*, 2003. Vol. 14, No. 6.

76. Pyysiäinen, I., Lindeman, M. & Honkela, T. Counterintuitiveness as the Hallmark of Religiosity. *Religion*. Vol. 33, No. 4, pp. 341-355.

77. Raike, A., Honkela, T., Jokinen, M. & Koskinen, L. CinemaSense Portal and Neural Network Analysis Method for Supporting Students of Linguistic Minorities in Distance Learning. *Proc. of Cumulus Conference*, Tallinn, Estonia, May 8-11, 2003, Cumulus Working Papers, University of Art and Design Helsinki, pp. 62-67.

78. Raiko, T., Valpola, H., Östman, T. & Karhunen, J. Missing Values in Hierarchical Nonlinear Factor Analysis. *Proc. of Joint 13th Int. Conf. on Artificial Neural Networks and 10th Int. Conf. on Neural Information Processing*, Istanbul, Turkey, June 26-29, 2003. pp. 185-189.

79. Raivio, K., Simula, O., Laiho, J. & Lehtimäki, P. Analysis of Mobile Radio Access Network Using the Self-Organizing Map. *Proc. of International Symposium on Integrated Network Management*, Colorado Springs, Colorado, USA, March 24-28, 2003. pp. 439-451.

80. Raju, K. & Ristaniemi, T. Exploiting Independence to Cancel Interferences due to Adjacent Cells in a DS-CDMA System. *Proc. of 13th IEEE Symposium on Personal Indoor Mobile Radio Communication (PIMRC 2003)*, Beijing, China, September 7-10, 2003.

81. Ristaniemi, T., Raju, K., Karhunen, J. & Oja, E. Inter-Cell Interference Cancellation in CDMA Array Systems by Independent Component Analysis. *Proc. of 4th Int. Symp. on Independent Component Analysis and Blind Signal Separation (ICA2003)*, Nara, Japan, April 2003. pp. 739-744.

82. Rummukainen, M., Laaksonen, J. & Koskela, M. An Efficiency Comparison of Two Content-Based Image Retrieval Systems, GIFT and PicSOM. *Proc. of International Conference on Image and Video Retrieval (CIVR 2003)*, Urbana, IL, USA, July 2003. pp. 500-509.

83. Salojärvi, J., Kaski, S. & Sinkkonen, J. Discriminative Clustering in Fisher Metrics. *Supplementary Proceedings ICANN/ICONIP 2003*, Istanbul, Turkey, June 26-29, 2003. pp. 161- 164.

84. Salojärvi, J., Kojo, I., Simola, J. & Kaski, S. Can Relevance be Inferred from Eye Movements in Information Retrieval? *Proc. of Workshop on Self-Organizing Maps (WSOM'03)*, Hibikino, Kitakyushu, Japan, September 11-14, 2003. pp. 261-266.

85. Salojärvi, J., Puolamäki, K. & Kaski, S. Relevance Feedback from Eye Movements for Proactive Information Retrieval. Espoo, Finland: Helsinki University of Technology, 2003. (Publications in Computer and Information Science Report A73).

86. Seppänen, J. K., Bingham, E. & Mannila, H. A Simple Algorithm for Topic Identification in 0-1 Data. *Knowledge Discovery in Databases: PKDD 2003, Proc. of 7th European Conference on Principles and Practice of Knowledge Discovery in Databases*, Cavtat-Dubrovnik, Croatia, September 22-26, 2003. pp. 423-434.

87. Siivola, V., Hirsimäki, T., Creutz, M. & Kurimo, M. Unlimited Vocabulary Speech Recognition Based on Morphs Discovered in an Unsupervised Manner. *Proc. of 8th European Conference on Speech Communication and Technology (Eurospeech)*, Geneva, Switzerland, September 1-4, 2003. pp. 2293-2296.

88. Siivola, V. & Honkela, A. A State-Space Method for Language Modeling. *Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding*, St. Thomas, US Virgin Islands, November 30 - December 4, 2003. pp. 548-553.

89. Sinkkonen, J., Nikkilä, J., Lahti, L. & Kaski, S. Associative Clustering by Maximizing a Bayes Factor. Espoo, Finland: Helsinki University of Technology, 2003. (Publications in Computer and Information Science Report A68).

90. Sirola, M. Conceptual Decision Model. *Proc. of IEEE International Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS'2003)*, Lviv, Ukraine, September 8-10, 2003. pp. 69-72.

91. Sirola, M. Decision Concepts. *Proc. of IEEE International Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS'2003)*, Lviv, Ukraine, September 8-10, 2003. pp. 59-62.

92. Sirola, M. Using Conceptual Decision Model in a Case Study. *Proc. of International Conference on Knowledge-Based Intelligent Information & Engineering Systems (KES'2003)*, Oxford, United Kingdom, September 3-5, 2003. pp. 126-133.

93. Sirola, M. Conceptual Decision Model. *International Journal of Computing*, 2003. Vol. 2, Issue 3, pp. 69-73.

94. Sjöberg, M., Laaksonen, J. & Viitaniemi, V. Using Image Segments in PicSOM CBIR System. *Proc. of 13th Scandinavian Conference on Image Analysis (SCIA 2003)*, Göteborg, Sweden, June 29-July 2, 2003. Springer, pp. 1106-1113.

95. Somervuo, P. Experiments with Linear and Nonlinear Feature Transformations in HMM Based Phone Recognition. *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Hong Kong, 2003. vol. 1, pp. 52-55.

96. Somervuo, P. Self-Organizing Map of Symbol Strings with Smooth Symbol Averaging. *Proc. of Workshop on Self-Organizing Maps (WSOM'03)*, Hibikino, Kitakyushu, Japan, September 11-14, 2003. (CD-ROM)

97. Somervuo, P. Speech Dimensionality Analysis on Hypercubical Self-Organizing Maps. *Neural Processing Letters*, 2003. Vol. 17, No. 2, pp. 125-136.

98. Somervuo, P., Chen, B. & Zhu, Q. Feature Transformations and Combinations for Improving ASR Performance. *Proc. of 8th European Conference on Speech Communication and Technology (Eurospeech)*, Geneva, Switzerland, September 1-4, 2003. Vol. I, pp. 477-480.

99. Somervuo, P. & Härmä, A. Analyzing Bird Song Syllables on the Self-Organizing Map. *Proc. of Workshop on Self-Organizing Maps (WSOM'03)*, Hibikino, Japan, September 11-14, 2003. (CD-ROM)

100. Sulkava, M. & Hollmén, J. Finding Profiles of Forest Nutrition by Clustering of the Self-Organizing Map. *Proc. of Workshop on Self-Organizing Maps (WSOM'03)*, Hibikino, Kitakyushu, Japan, September 11-14, 2003. pp. 243-248. (CD-ROM)

101. Särelä, J. & Vigário, R. A Bayesian Approach to Overlearning in ICA. Espoo, Finland: Helsinki University of Technology, 2003. (Publications in Computer and Information Science Report A70).

102. Valpola, H., Harva, M. & Karhunen, J. Hierarchical Models of Variance Sources. *4th Int. Symp. on Independent Component Analysis and Blind Signal Separation (ICA2003)*, Nara, Japan, April 2003. pp. 83-88.

103. Valpola, H., Honkela, A., Harva, M., Ilin, A., Raiko, T. & Östman, T. Bayes Blocks Software Library. Computer program. 2003.
http://www.cis.hut.fi/projects/bayes/software/

104. Valpola, H., Oja, E., Ilin, A., Honkela, A. & Karhunen, J. Nonlinear Blind Source Separation by Variational Bayesian Learning. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, 2003. Vol. E86-A, No. 3, pp. 532-541.

105. Valpola, H., Östman, T. & Karhunen, J. Nonlinear Independent Factor Analysis by Hierarchical Models. *Proc. of 4th Int. Symp. on Independent Component Analysis and Blind Signal Separation (ICA2003)*, Nara, Japan, April 2003. pp. 257-262.

106. Venna, J. & Kaski, S. Visualizing High-Dimensional Posterior Distributions in Bayesian Modeling. In: Kaynak, O., Alpaydin, E., Oja, E. & Xu, L. (eds.), *Artificial Neural Networks and Neural Information Processing - Supplementary Proc. ICANN/ICONIP 2003*, Istanbul, Turkey, June, pp. 165-168.

107. Venna, J., Kaski, S. & Peltonen, J. Visualizations for Assessing Convergence and Mixing of MCMC. In: Lavrac, N., Gamberger, D., Blockeel, H. & Todorovski, L. (eds.), *Proc. of 14th European Conference on Machine Learning (ECML 2003)*. Berlin 2003, Springer, pp. 432-443.

108. Vesanto, J. & Hollmén, J. An Automated Report Generation Tool for the Data Understanding Phase. In: Abraham, A. & Jain, L. (eds.), *Innovations in Intelligent Systems: Design, Management and Applications, Studies in Fuzziness and Soft Computing*. Heidelberg 2003, Springer (Physica) Verlag, chapter 5.

109. Vesanto, J., Sulkava, M. & Hollmén, J. On the Decomposition of the Self-Organizing Map Distortion Measure. *Proc. of Workshop on Self-Organizing Maps (WSOM'03)*, Hibikino, Kitakyushu, Japan, September 11-14, 2003. pp. 11-16.

110. Vigário, R., Ziehe, A., Müller, K.-L., Wübbeler, G., Trahms, L., Mackert, B.-M., Curio, G., Jousmäki, V., Särelä, J. & Oja, E. Blind Decomposition of Multimodal Evoked Responses and DC Fields. In: Sommer, F. & Wichert, A. (eds.), *Exploratory Analysis and Data Modeling in Functional Neuroimaging*. Cambridge, MA 2003, MIT Press, pp. 163-191.

II From Data to Knowledge Research Unit Research Projects under the CIS Laboratory

Chapter 17

From Data to Knowledge Research Unit

Heikki Mannila, Jaakko Hollmén, Ella Bingham, Johan Himberg, Mikko Koivisto, Anne Patrikainen, Salla Ruosaari, Jouni K. Seppänen, Mikko Katajamaa, Heli Juntunen, Nikolaj Tatti, Antti Rasinen, Kalle Korpiaho, Jaripekka Juhala, Antti Savolainen, Mikko Korpela, Janne Toivonen

17.1 Data mining

The work 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. One of the basic questions is how much does the collection of frequent patterns tell us about the underlying distribution. We have analyzed the use of maximum entropy approaches to inferring distributions from frequent pattern collections and obtained quite good empirical results [6]. Another major question is finding structure in large collection of 0-1 data: the results include a simple model of topics in 0-1 data, and simple algorithms for finding the topic structure [5]. In industrial cooperation projects we have recently developed simple and efficient algorithms for on-line clustering.

The combination of probabilistic and algorithmic techniques is also visible in several new themes. One major new theme in the work is in finding good segmentations for sequences. The (k,h)-segmentation problem and algorithms [2] show how one can locate recurrent sources from sequences; the approach applies to basically any probabilistic model for the generation of points in the sequences. We have also looked at the question of finding fragments of total orders from unordered data [1], which seems to be a fruitful approach. We are also investigating different approaches to subspace clustering.

On pure pattern discovery area, topics include approximation of frequent set collections and pattern discovery algorithms [3].

The work on combining local and global analysis in data mining will continue. Potential new themes include spectral clustering, interplay of probabilistic clustering and frequent sets [4], and word discovery from sequences. The work has lots of connections to applications, e.g., in paleontology and genomics.

References

[1] A. Gionis, T. Kujala and H. Mannila: Fragments of order. *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* P. Domingos et al. (eds.), pages 129–136, 2003.

[2] A. Gionis and H. Mannila: Finding recurrent sources in sequences. *The Seventh Annual International Conference on Research in Computational Molecular Biology*, W. Miller, M. Vingron, S. Istrail, P. Pevzner, M. Waterman (eds.), pages 123–130, ACM, 2003.

[3] D. Gunopulos, R. Khardon, H. Mannila, S. Saluja, H. Toivonen, and R.S. Sarma. Discovering all most specific sentences. *ACM Transactions on Database Systems* (28)2:140–174, 2003.

[4] Jaakko Hollmén, Jouni K. Seppänen, and Heikki Mannila. Mixture models and frequent sets: combining global and local methods for 0-1 data. In Daniel Barbará and Chandrika Kamath, editors, *Proceedings of the Third SIAM International Conference on Data Mining*, pages 289–293. Society of Industrial and Applied Mathematics, 2003.

[5] Jouni K. Seppänen, Ella Bingham, and Heikki Mannila. A simple algorithm for topic identification in 0-1 data. In Nada Lavrač, Dragan Gamberger, Ljupčo Todorovski, and Hendrik Blockeel, editors, *Knowledge Discovery in Databases: PKDD 2003. 7th European Conference on Principles and Practice of Knowledge Discovery*

in Databases. Cavtat-Dubrovnik, Croatia, September 2003, Proceedings, number 2838 in Lecture Notes in Artificial Intelligence, pages 423–434. Springer, 2003.

[6] D. Pavlov, H. Mannila, and P. Smyth. Beyond independence: probabilistic methods for query approximation on binary transaction data. *IEEE Transactions on Data and Knowledge Engineering* (15)6:1409–1421, 2003

17.2 Latent topics in 0-1 data

Large collections of 0–1 data occur in many applications, such as information retrieval, web browsing, telecommunications, and market basket analysis. While the dimensionality of such data sets can be large, the variables (or attributes) are seldom completely independent. Rather, it is natural to assume that the attributes are organized into (possibly overlapping) *topics*, i.e., collections of variables whose occurrences are somehow connected to each other. For example, in document data the topics correspond to topics of the document: e.g., phrases "data mining", "decision trees" and "association rules" probably are included in one topic, which might be called the "data mining" topic. In supermarket market basket data, the topics could correspond to classes of products such as soft drinks, vegetables, etc. In discretized gene expression data topics could correspond to groups of genes that are expressed in similar conditions or tissues.

Finding topics from data is by no means easy: the topics can be overlapping; a particular topic may be active only for a subset of documents; all attributes in a topic might not be present in the same observation. In the papers [1] and [2] we describe methods to estimate these hidden topics in 0-1 data. We specify several data models and give algorithms for finding the topics. An example of our topic model is given in Figure 17.1. The observed data are generated by interactions between independent latent topics: Each topic has a probability s_j of being active in an observation vector. The topics j then generate occurrences of variables $A, B, C, \ldots$ according to some topic-variable probabilities that are listed in matrix $\mathbf{A}$.

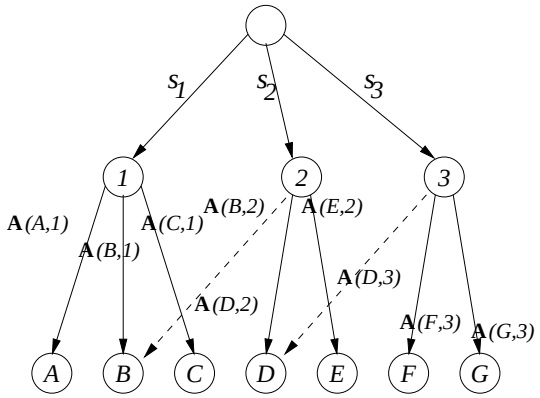


Figure 17.1: An example topic model. Topics 1, 2 and 3 are generated independently of each other with probabilities s_1, s_2 and s_3. The topics then generate observed variables with probabilities $\mathbf{A}(i, j)$. The dashed arrows indicate that a variable may be generated by several topics.

References

[1] Ella Bingham, Heikki Mannila, and Jouni K. Seppänen. Topics in 0-1 data. In David Hand, Daniel Keim, and Raymond Ng, editors, *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 450–455, Edmonton, Alberta, Canada, July 2002.

[2] Jouni K. Seppänen, Ella Bingham, and Heikki Mannila. A simple algorithm for topic identification in 0-1 data. In Nada Lavrač, Dragan Gamberger, Ljupčo Todorovski, and Hendrik Blockeel, editors, *Knowledge Discovery in Databases: PKDD*

2003. 7th European Conference on Principles and Practice of Knowledge Discovery in Databases. Cavtat-Dubrovnik, Croatia, September 2003, Proceedings, number 2838 in Lecture Notes in Artificial Intelligence, pages 423–434. Springer, 2003.

17.3 Applications in bioinformatics

Gene expression data analysis

Thousands of simultaneous gene expression measurements can be obtained with the microarray platform. As a result of an experiment, the analysts are faced with an abundance of data, usually a $N \times p$ matrix of continuous gene expression measurements, where N denotes the number of measured genes (usually in the thousands) and p the number of samples or subjects (usually a handful). In close collaboration with specialists in relevant fields in biology and medicine, we have analyzed this type of gene expression data in various cancer investigations [1,3,6,7], where the patient material is well characterized and other information is also available.

The analysis of the resulting gene expression data matrix can proceed in many alternative ways. We have used signal decomposition methods, for instance, principal component analysis [3,6,7] and non-negative matrix factorization [5] to yield meaningful components from the data. For instance, projections of data on the first principal component have been used as a score for collective difference in expression between the samples and the reference. In order to avoid stating random findings as true, we have extensively used the permutation testing in validating the results with the data set at hand [1,2,3,6,7]. Findings from the screening type of studies should be externally confirmed [1,3,6,7].

We have also examined publicly available gene expression data from baker's yeast [2]. Our statistical analysis indicates a correlation between genes located in the same chromosome that is only partially explained by known regulation mechanisms. These mechanisms function at a small spatial range, and indeed genes that are located close to each other are more tightly co-regulated; but also genes far away from each other show a small but significant correlation. By analyzing gene expression data in combination with other sources of data, one can make improved inferences.

Currently, the work continues with method development in the probabilistic framework to combine several sources of data, and to draw improved inferences based on the joint data set. The immediate application area is found in our collaborative cancer research: we are working on a project aiming at finding tumor markers of work-related lung cancers. Existing measurements include gene expression data from the microarray platform, copy number alteration measurements along the chromosome, characterization of the patient material, and gene annotation databases.

Quality control of microarray data

One strand of our work has investigated quality control of data originating from the microarray measurement platform, based on image analysis of scanned images of hybridized microarrays [4]. The goal is to be able to automatically classify spots into good and faulty spots, so that no erroneous spot would enter the subsequent analysis, therefore possibly causing bias in results. We extract features from the spot image describing shape, regularity and uniformity, and train a Naive Bayes classifier on the extracted features using a pre-labeled database of spot images. Furthermore, we describe a non-symmetric cost model in the cost-sensitive classification setting. Out of the three repetitions of the same measurement, we should allow as many good measurements as possible to enter to subsequent analysis, but to prevent an erroneous spot to be taken into account in the analysis phase. The results are assessed with Receiver Operating Characteristic (ROC) curves in the classification setting and expected costs in the cost-sensitive classification setting.

We plan to investigate the extension of these techniques to a three-color microarray platform, where a third channel is used to effectively bind to all probes of the array.

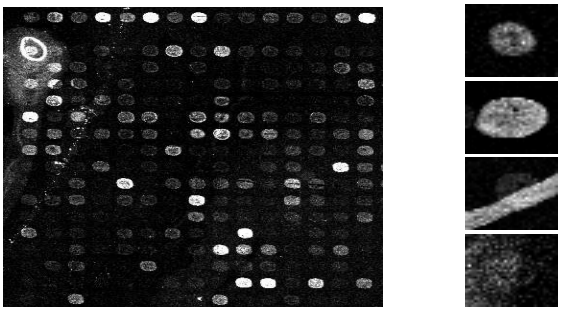


Figure 17.2: Example of an image of microarray containing numerous faulty spots is shown left. Four examples on the images on the right side demonstrate possible problems, for instance, spots of varying sizes in the two upper images and scratches and noise in the two lower images, respectively.

References

[1] Eeva Kettunen, Sisko Anttila, Jouni K. Seppänen, Antti Karjalainen, Henrik Edgren, Irmeli Lindström, Reijo Salovaara, Anna-Maria Nissén, Jarmo Salo, Karin Mattson, Jaakko Hollmén, Sakari Knuutila, and Harriet Wikman. 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*, In press.

[2] Heikki Mannila, Anne Patrikainen, Jouni K. Seppänen, and Juha Kere. Long-range control of expression in yeast. *Bioinformatics*, 18(3):482–483, 2002.

[3] Tarja Niini, Kim Vettenranta, Jaakko Hollmén, Marcelo L. Larramendy, Yan Aalto, Harriet Wikman, Bálint Nagy, Jouni K. Seppänen, Anna Ferrer Salvador, Heikki Mannila, Ulla M. Saarinen-Pihkala, and Sakari Knuutila. Expression of myeloid-specific genes in childhood acute lymphoblastic leukemia — a cDNA array study. *Leukemia*, 16(11):2213–2221, 2002. Nature Publishing Group.

[4] Salla Ruosaari and Jaakko Hollmén. Image analysis for detecting faulty spots from microarray images. In Steffen Lange, Ken Satoh, and Carl H. Smith, editors, *Proceedings of the 5th International Conference on Discovery Science (DS 2002)*, volume 2534 of *Lecture Notes in Computer Science*, pages 259–266. Springer-Verlag, 2002.

[5] Jouni K. Seppänen, Jaakko Hollmén, Ella Bingham, and Heikki Mannila. Nonnegative Matrix Factorization on Gene Expression Data. *Bioinformatics 2002*, Bergen, Norway, April 4-7, 2002. poster 49.

[6] Harriet Wikman, Eeva Kettunen, Jouni K. Seppänen, Antti Karjalainen, Jaakko Hollmén, Sisko Anttila, and Sakari Knuutila. Identification of differentially expressed genes in pulmonary adenocarcinoma by using a cDNA array. *Oncogene*, 21(37):5804–5813, 2002. Nature Publishing Group.

[7] Ying Zhu, Jaakko Hollmén, Riikka Räty, Yan Aalto, Balint Nagy, Erkki Elonen, Juha Kere, Heikki Mannila, Kaarle Franssila, and Sakari Knuutila. Investigatory and analytical approaches to differential gene expression profiling in mantle cell lymphoma. *British Journal of Haematology*, 119(4):905–915, 2002.

Publications of the From Data to Knowledge Research Unit

Publications are in alphabetical order by the first author.

2002

1. Bingham, E., Mannila, H. & Seppänen, J. K. Topics in 0-1 Data. *Proc. of 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Edmonton, Alberta, Canada, July 23-26, 2002. pp. 450-455.

2. Bykowski, A., Seppänen, J. & Hollmén, J. Model-independent Bounding of the Supports of Boolean Formulae in Binary Data. *Proc. of First International Workshop Knowledge Discovery in Inductive Databases (KDID'02)*, Helsinki, Finland, August 2002. pp. 20-31.

3. de Raedt, L., Jaeger, M., Lee, S. D. & Mannila, H. A Theory of Inductive Query Answering. *Proc. of 2002 IEEE International Conference on Data Mining (ICDE 2002)*, Maebashi City, Japan, December 9-12, 2002. pp. 123-130.

4. Elomaa, T., Mannila, H. & Toivonen, H. (eds.) *Proc. of ECML 2002 - 12th European Conference on Machine Learning, LNCS 2430*. Springer, 2002.

5. Elomaa, T., Mannila, H. & Toivonen, H. (eds.) *Proc. of Principles of Data Mining and Knowledge Discovery - 6th European Conference, PKDD 2002, LNCS 2431*. Springer, 2002.

6. Grossman, R., Han, J., Kumar, V., Mannila, H. & Motwani, R. (eds.) *Proceedings of the Second SIAM International Conference on Data Mining, SIAM 2002*; ISBN 0-89871-517-2. 2002.

7. Han, J., Altman, R. B., Kumar, V., Mannila, H. & Pregibon, D. Emerging Scientific Applications in Data Mining. *Communications of the ACM*, 2002. Vol. 45, No. 8, pp. 54-58.

8. Koivisto, M., Perola, M., Varilo, T., Hennah, W., Ekelund, J., Lukk, M., Peltonen, L., Ukkonen, E. & Mannila, H. An MDL Method for Finding Haplotype Blocks and for Estimating the Strength of Haplotype Block Boundaries. *Proc. of Pacific Symposium on Biocomputing 2003*, Kauai, Hawaii, January 3-7, 2003. 2002, World Scientific, pp. 502-513.

9. Leung, C.K., Ng, R. & Mannila, H. OSSM: A Segmentation Approach to Optimize Frequency Counting. *Proc. of 18th International Conference on Data Engineering (ICDE 2002)*, San Jose, CA, USA, February 26 - March 1, 2002. pp. 583-593.

10. Mannila, H. Global and Local Methods in Data Mining: Basic Techniques and Open Problems. *Proc. of 29th International Colloquium on Automata, Languages, and Programming, ICALP 2002*, Malaga, Spain, July 2002. 2002 , Springer-Verlag, pp. 57-68.

11. Mannila, H., Patrikainen, A., Seppänen, J. K. & Kere, J. Long-Range Control of Expression in Yeast. *Bioinformatics*, 2002. Vol. 18, pp. 482-483.

12. Niini, T., Vettenranta, K., Hollmén, J., Larramendy, M. L., Aalto, Y., Wikman, H., Nagy, B., Seppänen, J. K., Salvador, A. F., Mannila, H., Saarinen-Pihkala, U. M. & Knuutila, S. Expression of Myeloid-Specific Genes in Childhood Acute Lymphoblastic Leukemia – a cDNA Array Study. *Leukemia*, 2002. Vol. 16, No. 11, pp. 2213-2221.

13. Onkamo, P., Ollikainen, V., Sevon, P., Toivonen, H.T.T., Mannila, H. & Kere, J. Association Analysis for Quantitative Traits by Data Mining: QHPM. *The Annals of Human Genetics*, 2002. Vol. 66, pp. 419-429.

14. Ruosaari, S. & Hollmén, J. Image Analysis for Detecting Faulty Spots from Microarray Images. *Proc. of 5th International Conference on Discovery Science (DS'2002)*, Lübeck, Germany, November 24-26, 2002. Springer, pp. 259-266.

15. Salmenkivi, M., Kere, J. & Mannila, H. Genome Segmentation using Piecewise Constant Intensity Models and Reversible Jump MCMC. *Bioinformatics* (European Computational Biology Conference 2002), 2002. Vol. 18, Supplement 2, pp. S211-S218.

16. Seppänen, J. K., Hollmén, J., Bingham, E. & Mannila, H. Nonnegative Matrix Factorization on Gene Expression Data. *Bioinformatics 2002*, Bergen, Norway, April 4-7, 2002. poster 49. http://www.ii.uib.no/bio2002/abstractsbio2002.pdf

17. Wikman, H., Kettunen, E., Seppänen, J. K., Karjalainen, A., Hollmén, J., Anttila, S. & Knuutila, S. Identification of Differentially Expressed Genes in Pulmonary Adenocarcinoma by Using a cDNA Array. *Oncogene*, 2002. Vol. 21, No. 37, pp. 5804-5813.

18. Zhu, Y., Hollmén, J., Räty, R., Aalto, Y., Nagy, B., Elonen, E., Kere, J., Mannila, H., Franssila, K. & Knuutila, S. Investigatory and Analytical Approaches to Differential Gene Expression Profiling in Mantle Cell Lymphoma. *British Journal of Haematology*, 2002. Vol. 119, No. 4, pp. 905-915.

2003

1. Bykowski, A., Seppänen, J. & Hollmén, J. Model-Independent Bounding of the Supports of Boolean Formulae in Binary Data. In: Lanzi, P. & Meo, R. (eds.), *Database Support for Data Mining Applications*. Heidelberg, Springer-Verlag, in press.

2. Gionis, A., Kujala, T. & Mannila, H. Fragments of Order. *Proc. of ACM SIGKDD 2003*, Washington, D.C., USA, August 24-27, 2003. pp. 129-136.

3. Gionis, A. & Mannila, H. Finding Recurrent Sources in Sequences. *Proc. of ACM ReCOMB 2003*, Berlin, Germany, April 10-13, 2003. pp. 123-130.

4. Gunopulos, D., Khardon, R., Mannila, H., Saluja, S., Toivonen, H. & Sarma, R.S. Discovering All Most Specific Sentences. *ACM Transactions on Database Systems*, 2003. Vol. 28, No. 2, pp. 140-174.

5. Hollmén, J., Seppänen, J.K. & Mannila, H. Mixture Models and Frequent Sets: Combining Global and Local Methods for 0-1 Data. *Proc. of Third SIAM International Conference on Data Mining*, San Francisco, CA, USA, May 1-3, 2003. Society of Industrial and Applied Mathematics, pp. 289-293.

6. Katajamaa, M. & Hollmén, J. Simulation Model for Gene Expression Data. *Currents in Computational Molecular Biology 2003, Proc. of 7th Annual International Conference on Research in Computational Molecular Biology, RECOMB*, April 2003. pp. 249-250.

7. 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). *Cancer Genetics and Cytogenetics*, 2003. In press.

8. Leino, A., Mannila, H., Pitkänen, R.-L. Rule Discovery and Probabilistic Modeling for Onomastic Data. *Knowledge Discovery in Databases: PKDD 2003, Proc. of 7th European Conference on Principles and Practice of Knowledge Discovery in Databases*, Cavtat-Dubrovnik, Croatia, September 22-26, 2003. pp. 291-302.

9. Mielikäinen, T. & Mannila, H. The Pattern Ordering Problem. *Knowledge Discovery in Databases: PKDD 2003, Proc. of 7th European Conference on Principles and Practice of Knowledge Discovery in Databases*, Cavtat-Dubrovnik, Croatia, September 22-26, 2003. pp. 327-338.

10. Pavlov, D., Mannila, H. & Smyth, P. Beyond Independence: Probabilistic Methods for Query Approximation on Binary Transaction Data. *IEEE Transactions on Data and Knowledge Engineering*, 2003. Vol. 15, No. 6, pp. 1409-1421.

11. Ruosaari, S. & Hollmén, J. Identifying Differentially Expressed Genes. *Proc. of TICSP Workshop on Computational Systems Biology*, Tampere, Finland, June 16-17, 2003.

12. Ruosaari, S. & Hollmén, J. Identifying Differentially Expressed Genes with Bootstrap-Based Testing. *Proc. of Bioinformatics 2003*, Helsinki, Finland, May 22-24, 2003.

13. Seppänen, J. K., Bingham, E. & Mannila, H. A Simple Algorithm for Topic Identification in 0-1 Data. *Knowledge Discovery in Databases: PKDD 2003, Proc. of 7th European Conference on Principles and Practice of Knowledge Discovery in Databases*, Cavtat-Dubrovnik, Croatia, September 22-26, 2003. pp. 423-434.