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 some aspects of neural networks, 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, 15 post-doctoral researchers, 25 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 (FDK) research unit**, 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 2000 and 2001. 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 beginning of 2000, was thoroughly explained in the triennial reports 1994 - 1996 and 1997 - 1999. The web pages of the laboratory, `http://www.cis.hut.fi/` also contain up-to-date texts.

To briefly list the main achievements and highlights of the period 2000 - 2001, the laboratory produced 6 D.Sc. (Eng.) degrees, 4 Lic.Tech. degrees, and 38 M.Sc. (Eng.) degrees. The number of scientific publications appearing during the period was 194, of which 38 were journal papers. Special mention should be given to three scientific text-books authored or co-authored by our staff: *Self-Organizing Maps* (third, extended edition, T. Kohonen, Springer, 2001); *Independent Component Analysis* (A. Hyvärinen, J. Karhunen, and E. Oja, Wiley-Interscience, 2001); and *Principles of Data Mining* (D. Hand, H. Mannila, and P. Smyth, The MIT Press, 2001). Such text-books, with wide circulation, are indispensable in spreading our research results to a wide international audience, including practitioners and students.

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. The largest effort in this respect was the *Second International Workshop on Independent Component Analysis and Blind Signal Separation*, which was arranged by us on June 20 to 22, 2000 at the Hanasaari conference center, Espoo. During the reporting period, Professor Erkki Oja was the President of the European Neural Network Society. Also, some prices and honours, both national and international, were granted to members of our staff, the most prestigious one being the nomination to the rank of Academician of Professor (Emeritus) Teuvo Kohonen.

Erkki Oja

Academy Professor
Director,
Neural Networks
Research Centre

Olli Simula

Professor
Director,
Laboratory of Computer
and Information Science

Heikki Mannila

Professor
Vice Director,
From Data to Knowledge
Research Unit

Personnel

Employees during 2000 – 2001

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
Samuel Kaski, D.Sc. (Tech.) until July 31, 2001 (Acting Professor, E. Oja's chair)
Mikko Kurimo, D.Sc. (Tech.) from Aug. 1, 2001 (Acting Professor, E. Oja's chair)
Petteri Pajunen, D.Sc. (Tech.) (Acting Professor, J. Karhunen's chair)
Jaakko Hollmén, D.Sc. (Tech.) from Jan. 1, 2001 (Acting Professor, H. Mannila's chair)

Post-doc researchers

Adrian Flanagan, D.Phil. (until Dec. 31, 2000)
Aapo Hyvärinen, D.Phil.
Jukka Iivarinen, D.Sc. (Tech.)
Sirkka-Liisa Joutsiniemi, D.Med.Sc.
Samuel Kaski, D.Sc. (Tech.)
Mikko Kurimo, D.Sc. (Tech.)
Maija-Liisa Laakso, D.Med.Sc.
Jorma Laaksonen, D.Sc. (Tech.)
Krista Lagus, D.Sc. (Tech.)
Kai Puolamäki, D.Phil.
Kimmo Raivio, D.Sc. (Tech.)
Miki Sirola, D.Sc. (Tech.), laboratory engineer
Panu Somervuo, D.Sc. (Tech.)
Harri Valpola, D.Sc. (Tech.)
Ricardo Vigário, D.Sc. (Tech.)

Post-graduate researchers

Matti Aksela
Esa Alhoniemi
Ella Bingham
Mathias Creutz (from Oct. 15, 2001)
Razvan Cristescu (until Oct. 31, 2000)
Ramunas Girdziusas
Johan Himberg (until Feb. 29, 2000)

Antti Honkela
Patrik Hoyer
Jarmo Hurri
Alexander Ilin
Mika Inki
Kimmo Kiviluoto (until July 31, 2000)
Mikko Koivisto
Markus Koskela
Ville Könönen
Sampsa Laine
Simona Mălăroiu (until Aug. 30, 2000)
Janne Nikkilä
Jaakko Peltonen
Markus Peura (until March 18, 2001)
Karthikesh Raju
Jarkko Salojärvi
Jouni Seppänen
Vesa Siivola
Janne Sinkkonen
Jaakko Särelä
Jarkko Venna
Juha Vesanto
Vuokko Vuori

Under-graduate researchers

Jussi Ahola (until Dec. 31, 2001)
Juha Aulio
Hugo Gävert (until May 14, 2001)
Teemu Hirsimäki
Mikko Katajamaa
Arto Klami
Kalle Korpiaho
Jukka Kuusisto
Sami Laakso (until March 31, 2001)
Pasi Lehtimäki
Teemu Murtola
Merja Oja
Vesa Paatero
Jussi Pakkanen
Jukka Parviainen
Anne Patrikainen
Tapani Raiko
Antti Rasinen
Toni Saarela
Timo Similä
Markus Siponen (until Dec. 31, 2000)
Harri Sulkava
Johanna Tikanmäki
Jarkko Tikka

Sampo Viiperi
Ville Viitaniemi

Support staff

Leila Koivisto, department secretary
Esko Kokkonen, maintenance assistant
Tarja Pihamaa, laboratory secretary
Teemu Pohjolainen, maintenance assistant
Markku Ranta, B.Eng., works engineer
Ville Rantamaula, maintenance assistant

Awards and activities

Prizes and scientific honours received by researchers of the unit

Emeritus Professor Teuvo Kohonen:

- Honorary title of Academician. 2000.
- E.R. Caianiello Award, Italy (SIREN - Italian Association for Neural Networks). 2000.
- IEEE Millennium Medal, USA. 2000.

Academy Professor Erkki Oja:

- IEEE Fellow, USA. 2000.
- Nokia Educational Award, Nokia Foundation. 2001.

Professor Olli Simula:

- IEEE Millennium Medal, USA. 2000.

Professor Juha Karhunen:

- Nokia Educational Award, Nokia Foundation. 2001.

Professor Jaakko Hollmén:

- Dissertation Award of the year 2000, The Finnish Society for Computer Science. 2001.

Dr. Aapo Hyvärinen:

- Nokia Educational Award, Nokia Foundation. 2001.

Dr. Miki Sirola:

- Best Paper Award at 14th Int. Conf. on Systems Science, Wroclaw, Poland, Sept. 11-14, 2001.

Important international positions of trust held by researchers of the unit

Academy Professor Erkki Oja:

- Editorial Board Member:
  IEEE Transactions on Neural Networks (USA) (until 2000)
  Neural Computation (USA)
  Pattern Recognition Letters (The Netherlands)
  Computational Intelligence - CITE (The Netherlands)
  International Journal of Pattern Recognition and and Artificial Intelligence (Singapore)
  International Journal of Computer Research (USA)
  Journal of Natural Computing (The Netherlands).
- International Association for Pattern Recognition (IAPR), USA. Member of the Governing Board, until 2000.
- European Neural Network Society (ENNS), Great Britain. President.
- Advisory Board of the Leiden Centre for Natural Computing, The Netherlands. Expert.
- Advisory Board of Springer Verlag book series on Natural Computing, The Netherlands. Expert.
- Evaluator in filling the academic chairs of professors, Kungliga Tekniska Högskolan, Sweden. 6 chairs: computer systems, human computer interaction, neuroinformatics (2), computer science (2). 2000.
- Evaluator in filling the academic chair of professor, computer science, University of Stockholm, 2001.
- Opponent at the doctoral dissertation of Alexander Ypma, Delft University of Technology, The Netherlands, 2001.
- Plenary talk "Independent components for interpretation and prediction." Symposium on Engineering and Intelligent Systems, Paisley, Scotland, June 27-30, 2000.
- Plenary talk "Machine learning in data mining." 10th Int. PEG Conference (PEG 2001), Tampere, Finland, June 23-26, 2001.
- Invited talk "Analyzing low-level visual features using content-based image retrieval." 7th Int. Conf. on Neural Information Processing, ICONIP2000, Taejon, Korea, Nov. 14-18, 2000.
- Invited talk "Application of adaptive committee classifiers in on-line character recognition." 2nd Int. Conf. on Advances in Pattern Recognition (ICAPR'01), Rio de Janeiro, Brazil, March 11-14, 2001.
- Plenary talk, "Unsupervised learning for data and signal analysis." ISI-2001, Dubai, United Arab Emirates, March 18-22, 2001.
- Tutorial, "Independent component analysis." SCIA-2001, Bergen, Norway, June 11-13, 2001.

- Plenary talk, "Independent component analysis." Norsig 2001, Trondheim, Norway, Oct. 19-20, 2001.
- Plenary talk, "Finding structure in signals, images, and data." 8th Int. Conf. on Neural Information Processing (ICONIP2001), Shanghai, China, Nov. 14-18, 2001.
- Conference Chair and Organizer, 2nd Int. Workshop on Independent Component Analysis and Blind Signal Separation, Helsinki, Finland, June 19-22, 2000
- Honorary Chair, EIS2000, Symposium on Engineering and Intelligent Systems, Paisley, Scotland, June 27-30, 2000.
- Program Co-Chair, ICONIP2000, Int. Conf. on Neural Information Processing, Taejon, Korea, Nov. 14-18, 2000.
- Program Committee Member, Tutorial Chairman, IJCNN2000, Int. Joint Conference on Neural Networks, Como, Italy, July 24-27, 2000.
- Program Committee Member:
  EUSIPCO2000, European Signal Processing Conference, Tampere, Finland, Sept. 4-8, 2000
  IEEE Int. Workshop on Biologically Motivated Computer Vision, Seoul, Korea, May 15-17, 2000
  Int. Conf. on Artificial Neural Networks (ICANN'01), Vienna, Austria, Aug. 21-25, 2001
  3rd Int. Conf. on Independent Component Analysis and Signal Separation (ICA2001), San Diego, Calif., USA, Dec. 9-13, 2001
  Int. Conf. on Neural Information Processing (ICONIP2001), Shanghai, China, Nov. 14-18, 2001
  15th European Conf. on Circuit Theory and Design, Espoo, Finland, Aug. 28-31, 2001
  Int. Conf. on Computational Intelligence for Modelling, Control and Automation, Las Vegas, USA, July 9-11, 2001
  Workshop on Self-Organizing Maps, Lincoln, UK, June 13-15, 2001.

Academician Teuvo Kohonen:

- Editorial Board Member:
  Neural Computation (USA)
  Connection Science (Great Britain)
  Journal of Intelligent Systems (Great Britain).
- Editor, Springer Series in Information Sciences.
- Honorary Editor, International Journal of Computer Research (USA).
- Evaluator in filling the academic chair of Professor, University of Texas at Austin, Computer Science Department, 2001.
- Plenary talk "New lines in the study of self-organizing maps." IIZUKA 2000, 6th Int. Conf. on Soft Computing, Iizuka, Japan, Oct. 1-4, 2000.
- Plenary talk "Self-organizing maps of massive document collections." Int. Joint Conf. on Neural Networks, IJCNN2000, Como, Italy, July 24-27, 2000.

- Plenary talk "A look into the self-organizing maps." 2nd Int. ICSC Symposium on Neural Computation, NC2000, Berlin, Germany, May 23-26, 2000.
- Plenary talk "An unsupervised learning method that produces organized representation from real information." ANNIMAB-1 Conference, Gothenburg, Sweden, May 13-16, 2000.
- Plenary talk "A projection tool for exploration of large datbases." 7th Int. Workshop PASE2000, Leuven, Belgium, April 2-5, 2000.
- Tutorial "Self-organizing maps." EUSIPCO2000, European Signal Processing Conference, Tampere, Finland, Sept. 4-8, 2000.
- Honorary Chairman, IIZUKA 2000, 6th Int. Conf. on Soft Computing, Iizuka, Japan, Oct. 1-4, 2000.
- Special guest lecture "The SOM for large amounts of symbolic data." Workshop on Self-Organizing Maps (WSOM'01), Lincoln, UK, June 13-15, 2001.
- Plenary talk "Mining and mapping biochemical data from the Internet." Int. Joint Conf. on Neural Networks (IJCNN'01), Washington D.C., USA, July 15-19, 2001.
- Plenary talk "Evolution of ideas for self-organizing neural networks." Asian Pasific Symposium on Life Science and Systems Engineering, Kitakyushu, Japan, July 25-26, 2001.
- Plenary talk "Self-organizing neural networks." European Conference on Circuit Theory and Design, Espoo, Finland, Aug. 28-31, 2001.

Professor Juha Karhunen:

- Editorial Board Member:
  Neural Processing Letters (The Netherlands)
  Neurocomputing (USA).
- Editor of the Proceedings of the 2nd International Workshop on Independent Component Analysis and Blind Signal Separation, Helsinki, Finland, June 19-22, 2000.
- Invited talk "Blind separation methods for CDMA Communications." Workshop on Machines that Learn, Snowbird, Utah, April 4-7, 2000.
- Invited talk "Nonlinear static and dynamic blind source separation using ensemble learning." Int. Joint Conf. on Neural Networks (IJCNN'01), Washington D.C., USA, July 15-19, 2001.
- Invited lecture "Nonlinear blind source separation and independent component analysis", German national graduate school at the University of Regensburg, Regensburg, Germany, Nov. 16, 2001.
- Program Committee Member:
  2nd International Workshop on Independent Component Analysis and Blind Signal Separation, Helsinki, Finland, June 19-22, 2000
  EUSIPCO2000, European Signal Processing Conference, Tampere, Finland, Sept. 4-8, 2000
  Int. Conf. on Advances in Intelligent Systems: Theory and Applications,

AISTA2000, Canberra, Australia, Feb. 2000
European Symposium on Neural Networks, ESANN2000, Bruges, Belgium, April 26-28, 2000
Int. Conf. on Artificial Neural Networks in Medicine and Biology, ANNIMAB-1, Gothenburg, Sweden, May 13-16, 2000
European Symposium on Artificial Neural Networks (ESANN2001), Bruges, Belgium, April 25-27, 2001.

- Session Chairman and Program Committee Member:
  Int. Workshop on Artificial Neural Networks (IWANN'01), Granada, Spain, June 13-15, 2001
  3rd Int. Conf. on Independent Component Analysis and Signal Separation (ICA2001), San Diego, Calif., USA, Dec. 9-13, 2001.

Professor Olli Simula:

- Editorial Board Member:
  International Journal of Knowledge-Based Intelligent Engineering Systems (Australia).
- European Neural Network Society (ENNS), Great Britain. Governing Board Member.
- IEEE - Computer Chapter (C-16), Finland. Chairman.
- Invited talk "Models from data analysis of industrial processes and telecommunication systems." Automazione'2001, Ancona, Italy, Nov. 22-23, 2001.
- Program Committee Member, Session Chairman, EUSIPCO2000, European Signal Processing Conference, Tampere, Finland, Sept. 4-8, 2000.
- Conference Co-Chair, European Conference on Circuit Theory and Design, Espoo, Finland, Aug. 28-31, 2001.
- Session Chairman, Program Committee Member, Int. Conf. on Artificial Neural Networks, Vienna, Austria, Aug. 21-25, 2001.
- Program Committee Member, URSI XXVI Convention on Radio Science, Tampere, Finland, Oct. 23-24, 2001.

Professor Heikki Mannila:

- Editor-in-Chief, Data Mining and Knowledge Discovery.
- Associate Editor, ACM Transactions on Internet Technology, 2001–.
- Action Editor, Journal of Machine Learning Research, 2001–.
- Member of Technical Advisory Board, Verity Inc.
- SIGKDD Chair's Advisory Council, –2000.
- SIGKDD Awards Chair, 2000.
- PKDD Steering Committee, 2001–.

- Program Committee Member:
  4th European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD 2000)
  6th SIGKDD Conf. on Data Mining and Knowledge Discovery (KDD'00)
  20th ACM SIGACT-SIGART-SIGMOD Symposium on Principles of Database and Knowledgebase Systems (PODS'01)
  7th SIGKDD Conf. on Data Mining and Knowledge Discovery (KDD'01)
  18th Int. Conf. on Machine Learning (ICML-2001) (Area Chair)
  17th Int. Joint Conf. on Artificial Intelligence (IJCAI'01)
  5th European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD 2001).

Professor Jaakko Hollmén:

- Invited talk "Self-organizing map: Introduction to methods and applications." Finnish-Taiwanese Workshop on Digital Communications, Espoo, Finland, May 14, 2001.

Dr. Samuel Kaski:

- Plenary talk "Learning metrics for exploratory data analysis." 2001 IEEE Signal Processing Society Workshop, Neural Networks for Signal Processing XI, Falmouth, Massach., USA, Sept. 10-12, 2001.

- Invited talk "SOM-based exploratory analysis of gene expression data." Workshop on Self-Organizing Maps (WSOM'01), Lincoln, UK, June 13-15, 2001.

- Session Chairman:
  2nd Int. Conf. on Intelligent Data Engineering and Automated Learning, Hong Kong, Dec. 13-15, 2000
  The 2001 IEEE Workshop on Neural Networks for Signal Processing, Falmouth, MA, USA, Sept. 10-12, 2001.

- Session Chairman and Program Committee Member, Workshop on Self-Organizing Maps (WSOM'01), Lincoln, UK, June 13-15, 2001.

- Program Committee Member:
  Int. Symposium on Intelligent Data Engineering and Automated Learning, IDEAL 2000, Hong Kong, Dec. 13-15, 2000
  Finnish Conf. on Artificial Intelligence, STeP 2000, Espoo, Finland, Aug. 28-31, 2000
  2nd Int. Conf. on Data Warehousing and Knowledge Discovery, DaWak 2000, Greenwich, UK, Sept. 4-6, 2000
  Int. Conf. on Artificial Neural Networks in Medicine and Biology, ANNIMAB-1, Gothenburg, Sweden, May 13-16, 2000
  Third Int. Conf. on Data Warehousing and Knowledge Discovery, Munich, Germany, Sept. 5-7, 2001.

Dr. Mikko Kurimo:

- Tutorial "Automatic reoognition of auditory speech." EUSIPCO'2000, Tampere, Finland, Sept. 5-8, 2000.

- Scientific Committee Member, 7th European Conference on Speech Communication and Technology, Eurospeech 2001, Aalborg, Denmark, Sept. 3-7, 2001.

Dr. Aapo Hyvärinen:

- Editor-in-Charge, International Journal of Neural Systems (Singapore).
- Invited talk ”Complexity pursuit: Using both nongaussianity and autocorrelations for signal separation.” Int. Conf on Neural Information Processing, Taejon, Korea, Nov. 14-18, 2000.
- Invited talk ”Advances in independent component analysis.” RIKEN Brain Science Institute seminar, Japan, Nov. 9, 2000
- Panel in Int. Joint Conf. on Neural Networks, Como, Italy, July 24-27, 2000. Panel topic: “Unsupervised learning, statistics, and signal separation.”
- Tutorial ”Independent component analysis, blind source separation, and natural image statistics.” The Copenhagen Summer School in Computer Vision, Denmark, Sept. 4-8, 2000.
- Invited talk “Independent component analysis and projection pursuit.” Int. Meeting of the Psychometric Society, Osaka, Japan, July 15-19, 2001.
- Invited talk “Independent component analysis: A tutorial.” Int. Conf. on Artificial Neural Networks, Vienna, Austria, August 12-14, 2001.
- Publicity Chair, Int. Workshop on Independent Component Analysis and Blind Source Separation (ICA 2000), Helsinki, Finland, June 19-22, 2000.
- Session Chair, Int. Workshop on Independent Component Analysis and Blind Source Separation (ICA 2000), Helsinki, Finland, June 19-22, 2000.
- Main Organizer and Chair of Special Session “Models of natural image statistics.” Int. Conf. on Neural Information Processing, Taejon, Korea, Nov. 14-18, 2000.
- Program Committee Member, Int. Workshop on Independent Component Analysis and Blind Source Separation, San Diego, CA, USA, Dec. 9-12, 2001.

Dr. Miki Sirola:

- IASTED, Technical Committee on Modelling and Simulation, Canada. Member.
- Session Chair:
IASTED Int. Conf. on Modelling, Identification and Control, Innsbruck, Austria, Feb. 14-17, 2000
IASTED Int. Conf. on Modelling, Identification and Control, Innsbruck, Austria, Feb. 19-22, 2001.
- Program Committee Member:
IASTED Int. Conf. on Control and Applications, Cancun, Mexico, May 24-27, 2000
IASTED Int. Conf. on Control and Applications, Banff, Canada, June 27-29, 2001.

Dr. Jukka Iivarinen:

- Treasurer, Int. Workshop on Independent Component Analysis and Blind Source Separation (ICA 2000), Helsinki, Finland, June 19-22, 2000.
- Session Chair and Program Committee Member, Finnish Conf. on Artificial Intelligence, STeP 2000, Espoo, Finland, Aug. 28-31, 2000.
- Session Chair, IASTED Int. Conf. on Visualization, Imaging, and Image Processing, Marbella, Spain, Sept. 3-5, 2001.

Dr. Krista Lagus:

- Invited talk, ”Studying similarities in term usage with self-organizing maps.” NordTerm 2001, Tuusula, Finland, June 13-16, 2001.

M.Sc. Antti Honkela:

- Invited talk ”An ensemble learning approach to nonlinear independent component analysis.” European Conf. on Circuit Theory and Design (ECCTD'01), Espoo, Finland, August 28-31, 2001.

M.Sc. Patrik Hoyer:

- Invited talk “Topographic maps from sparse coding.” CNS'2001 Workshop ‘Modeling cortical map development’ organized by J.A. Bednar and R. Miikkulainen, San Francisco, Calif., USA, July 5, 2001.

Important domestic positions of trust held by researchers of the unit

Academy Professor Erkki Oja:

- The Finnish Academy of Science and Letters. The group of mathematics and computer science. Vice Chairman.
- Nokia Foundation. Member of the Governing Board.
- Alfred Kordelin Foundation. Member of the Governing Board.
- Opponent at the doctoral dissertation of Samuli Visuri, HUT, 2001.

Academician Teuvo Kohonen:

- Studia Generalia lecture ”Can machines learn autonomously.” University of Helsinki, 2000.

Professor Olli Simula:

- Evaluator in filling the academic chairs of professors, University of Jyväskylä. 2 chairs on telecommunications, 2001.

Professor Juha Karhunen:

- Opponent at the doctoral dissertation of Harri Saarnisaari, University of Oulu, 2000.

- Invited talk "Riippumattomien komponenttien analyysi ja informaatioteoreettinen oppiminen" (in Finnish). The 25th anniversary seminar of the Pattern Recognition Society of Finland at Mekrijärvi, Ilomantsi, Finland, October 22, 2001.

Professor Jaakko Hollmén:

- Invited talk "Käyttäjien profilointi ja luokittelu väärinkäytösten paljastamiseksi matkapuhelinverkossa" (in Finnish). The Computer Science Days, Turku, Finland, May 21-22, 2001.

Dr. Samuel Kaski:

- Pattern Recognition Society of Finland. Member of the Governing Board.
- Finnish Artificial Intelligence Society. Member of the Governing Board.
- Opponent at the doctoral dissertation of Jani Lahnajärvi, Tampere University of Technology, 2001.

Dr. Jukka Iivarinen:

- Pattern Recognition Society of Finland. Chairman.

Dr. Jorma Laaksonen:

- Opponent at the doctoral dissertation of Marius Tico, Tampere University of Technology, 2001.

Dr. Krista Lagus:

- Finnish Cognitive Linguistics Association. Member of the Governing Board.

Other activities

Academy Professor Erkki Oja:

- interview for the science series program Prisma, TV1, in April, 2000.
- interview for Akatemia Apropos -journal No. 1/2000.

Professor Olli Simula:

- interview on subjects "International education" and "Relationships between industry/business and higher education" for International French Radio, FRI, program Magazine multimédia, in Paris, France, on April 27, 2001.

Professor Samuel Kaski:

- interview on subject "bioinformatics" for the science series program Prisma, TV1, on Nov. 22, 2000.

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

- Academy Professor Erkki Oja, Université Paris, I - Pantheon Sorbonne, France, April–May 2001.
- Dr. Mikko Kurimo, IDIAP, Switzerland, Jan.–May 2000.
- Dr. Jaakko Hollmén, Siemens Corporation Research and Development, Germany, Jan.–Aug. 2000.
- M.Sc. Vesa Siivola, IDIAP, Switzerland, Jan.–March 2000.
- Dr. Ricardo Vigário, GMD, Germany, Oct. 2000–Dec. 2001.
- Dr. Aapo Hyvärinen, RIKEN Brain Science Institute, Japan, Nov. 2000 (2 wks), April–May 2001 (6 wks), and Nov.–Dec. 2001 (2 wks).
- Lic.Sc. Razvan Cristescu, EPFL, Switzerland, Jan.–Oct. 2000.
- M.Sc. Tapani Raiko, Albert-Ludwigs-Universität, Freiburg, Germany, Oct. 2001–
- Dr. Aapo Hyvärinen, Technische Universität Berlin, Germany, Oct. 2001.

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

- Dr. Adrian Flanagan, EPFL, Switzerland, Jan.–Dec. 2000.
- Lic.Sc. Simona Mălăroiu, University of Bucharest, Romania, Jan.–Aug. 2000.
- M.Sc. Karthikesh Raju, Universität Ulm, Germany, Jan.–March 2000.
- Prof. Mark Girolami, University of Paisley, Scotland, June–Nov. 2000.
- Dr. Ata Kaban, University of Paisley, Scotland, June–Nov. 2000.
- M.Sc. Xavier Giannakopoulos, IDSIA, Switzerland, Feb.–March 2000.
- M.Sc. Artur Bykowski, INSA - Lyon, France, Jan.–June 2001
- M.Sc. Maria Funaro, University of Salerno, Italy, Feb.–June 2001.
- Dr. Diane Litman, AT&T Research Laboratories, USA, March–May, 2001 (8 wks).
- M.Sc. Hervé Le Borgne, National Polytechnical Institute of Grenoble, Signal and Image Laboratory, France, Sept.–Oct. 2001 (6 wks).
- Dr. Ignacio Diaz, Universidad de Oviedo, Gijon, Spain, June–Dec. 2001.

Courses

Courses given by the Laboratory of Computer and Information Science.

Spring Semester 2000

Code	Course	Lecturer	Course Assistant
Tik-61.140	Signal Processing Systems	O. Simula	M. Koskela, J. Parviainen
Tik-61.152	Seminar on Computer and Information Science	S. Kaski	
Tik-61.233	Computer Vision	J. Laaksonen	J. Iivarinen
Tik-61.256	Learning Models and Methods	P. Pajunen	E. Bingham
Tik-61.261	Principles of Neural Computing	J. Karhunen	V. Vuori
Tik-61.182	Special Course II: *Information Retrieval and Statistical Natural Language Processing*	S. Kaski, E. Oja, K. Lagus	
Tik-61.183	Special Course III: *Randomized Algorithms*	H. Mannila, J. Seppänen	

Fall Semester 2000

Code	Course	Lecturer	Course Assistant
Tik-61.123	Computer Architecture	M. Huttunen	M. Rättö
Tik-61.124	Special Project in Computer Architecture	M. Huttunen	M. Rättö
Tik-61.231	Principles of Pattern Recognition	V. Vuori	M. Aksela
Tik-61.238	Statistical Signal Modelling	P. Pajunen	E. Bingham
Tik-61.246	Digital Signal Processing and Filtering	O. Simula	M. Koskela, V. Siivola, J. Parviainen
Tik-61.247	Digital Image Processing	M. Peura	J. Särelä
Tik-61.263	Advanced Course in Neural Computing	S. Kaski	J. Peltonen
Tik-61.181	Special Course I: *Bioinformatics*	S. Kaski, H. Mannila	J. Nikkilä
Tik-61.184	Special Course IV: *Unsupervised Learning*	M. Girolami, A. Hyvärinen	P. Hoyer
Tik-122.101	Special Course V: *Data Mining and Telecommunications*	H. Mannila, J. Hollmén, J. Seppänen	

Spring Semester 2001

Code	Course	Lecturer	Course Assistant
Tik-61.140	Signal Processing Systems	O. Simula	M. Koskela, J. Parviainen
Tik-61.152	Seminar on Computer and Information Science	S. Kaski	
Tik-61.233	Computer Vision	J. Laaksonen	J. Iivarinen
Tik-61.256	Learning Models and Methods	P. Pajunen	E. Bingham
Tik-61.261	Principles of Neural Computing	J. Karhunen	V. Vuori
Tik-61.182	Special Course II: *Information Visualisation*	S. Kaski, K. Puolamäki	
Tik-122.102	Special Course VI: *Principles of Data Mining*	J. Hollmén, H. Mannila, J. Seppänen	

Fall Semester 2001

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. Aksela
T-61.238	Statistical Signal Modelling	P. Pajunen	E. Bingham
T-61.246	Digital Signal Processing and Filtering	O. Simula	M. Koskela, V. Siivola, J. Parviainen
T-61.247	Digital Image Processing	J. Laaksonen	J. Iivarinen
T-61.263	Advanced Course in Neural Computing	S. Kaski	J. Peltonen
T-61.271	Information Visualisation	K. Puolamäki	J. Venna
T-61.181	Special Course I: *Independent Component Analysis*	E. Oja, J. Karhunen	A. Honkela
T-61.184	Special Course IV: *Statistical (Adaptive) Language Modeling*	K. Lagus, M. Kurimo	
T-122.101	Special Course V: *Inference and Learning in Bayes Networks*	J. Hollmén, H. Mannila, J. Seppänen	

Doctoral dissertations

Bayesian ensemble learning for nonlinear factor analysis

Harri Valpola

Dissertation for the degree of Doctor of Science in Technology on 10 November 2000.

External examiners:
Mark Girolami (University of Paisley)
Petri Myllymäki (University of Helsinki)
Opponent:
Zoubin Ghahramani (University College London)

Abstract:
An active research topic in machine learning is the development of model structures which would be rich enough to represent relevant aspects of the observations but would still allow efficient learning and inference.

Linear factor analysis and related methods such as principal component analysis and independent component analysis are widely used feature extraction and data analysis techniques. They are computationally efficient but are restricted to linear models. Many natural phenomena are nonlinear and therefore several attempts have been made to generalise the model by relaxing the linearity assumption. The suggested approaches have suffered from overfitting and the computational complexity of many of the algorithms scales exponentially with respect to the number of factors, which makes the application of these methods to high dimensional factor spaces infeasible.

This thesis describes the development of a nonlinear extension of factor analysis. The learning algorithm is based on Bayesian probability theory and solves many of the problems related to overfitting. The unknown nonlinear generative mapping is modelled by a multi-layer perceptron network. The computational complexity of the algorithm scales quadratically with respect to the dimension of the factor space which makes it possible to use a significantly larger number of factors than with the previous algorithms. The feasibility of the algorithm is demonstrated in experiments with artificial and natural data sets. Extensions which combine the nonlinear model with non-Gaussian and dynamic models for the factors are introduced.

Text mining with the WEBSOM

Krista Lagus

Dissertation for the degree of Doctor of Science in Technology on 11 December 2000.

External examiners:
Dieter Merkl (Technische Universität Wien)
Pasi Koikkalainen (University of Jyväskylä)
Opponent:
Risto Miikkulainen (University of Texas at Austin)

Abstract:
The emerging field of text mining applies methods from data mining and exploratory data analysis to analyzing text collections and to conveying information to the user in an intuitive manner. Visual, map-like displays provide a powerful and fast medium for portraying information about large collections of text. Relationships between text items and collections, such as similarity, clusters, gaps and outliers can be communicated naturally using spatial relationships, shading, and colors.

In the WEBSOM method the self-organizing map (SOM) algorithm is used to automatically organize very large and high-dimensional collections of text documents onto two-dimensional map displays. The map forms a document landscape where similar documents appear close to each other at points of the regular map grid. The landscape can be labeled with automatically identified descriptive words that convey properties of each area and also act as landmarks during exploration. With the help of an HTML-based interactive tool the ordered landscape can be used in browsing the document collection and in performing searches on the map.

An organized map offers an overview of an unknown document collection helping the user in familiarizing herself with the domain. Map displays that are already familiar can be used as visual frames of reference for conveying properties of unknown text items. Static, thematically arranged document landscapes provide meaningful backgrounds for dynamic visualizations of for example time-related properties of the data. Search results can be visualized in the context of related documents.

Experiments on document collections of various sizes, text types, and languages show that the WEBSOM method is scalable and generally applicable. Preliminary results in a text retrieval experiment indicate that even when the additional value provided by the visualization is disregarded the document maps perform at least comparably with more conventional retrieval methods.

Self-organizing maps for signal and symbol sequences

Panu Somervuo

Dissertation for the degree of Doctor of Science in Technology on 17 December 2000.

External examiners:
Olli Ventä (Technical Research Centre of Finland)
Jukka Heikkonen (Helsinki University of Technology)
Opponent:
Kari Torkkola (Motorola, USA)

Abstract:
Temporal sequences arise from various kinds of sources in the nature. Sensory elements transform the events into measurements and corresponding feature vectors. Methods for comparing sequences are needed in retrieval, error correction, and recognition tasks. Due to the durational differences in the feature sequences and the variation and noise in the feature vectors, both temporal and spatial fluctuations must be tolerated in the comparison.

The topic of this thesis is the application of the Self-Organizing Map (SOM) and Learning Vector Quantization (LVQ) algorithms to the processing of data sequences with numeric or symbolic elements. Usually the models of the data are fixed-dimensional feature vectors in the SOM and LVQ. In this work also other models have been experimented such as hidden Markov models, feature vector sequences, and symbol sequences. The goal has been to utilize the principles of the unsupervised learning of the SOM and the error corrective learning of the LVQ while also taking the sequential nature of the input data into account.

The developed methods have been applied to the unsupervised segmentation of speech, forming of templates of variable-length feature vector sequences, improvement of the performance of a speech recognizer based on hidden Markov models, construction of a data-driven pronunciation dictionary for speech recognition, and clustering of all currently known protein sequences.

User profiling and classification for fraud detection in mobile communications networks

Jaakko Hollmén

Dissertation for the degree of Doctor of Science in Technology on 19 December 2000.

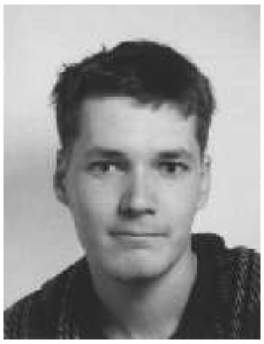
External examiners:
Jyrki Joutsensalo (University of Jyväskylä)
Henry Tirri (University of Helsinki)
Opponent:
Klaus Obermayer (Technische Universität Berlin)

Abstract:
The topic of this thesis is fraud detection in mobile communications networks by means of user profiling and classification techniques. The goal is to first identify relevant user groups based on call data and then to assign a user to a relevant group. Fraud may be defined as a dishonest or illegal use of services, with the intention to avoid service charges. Fraud detection is an important application, since network operators lose a relevant portion of their revenue to fraud. Whereas the intentions of the mobile phone users cannot be observed, it is assumed that the intentions are reflected in the call data. The call data is subsequently used in describing behavioral patterns of users. Neural networks and probabilistic models are employed in learning these usage patterns from call data. These models are used either to detect abrupt changes in established usage patterns or to recognize typical usage patterns of fraud. The methods are shown to be effective in detecting fraudulent behavior by empirically testing the methods with data from real mobile communications networks.

Attribute trees as adaptive object models in image analysis

Markus Peura

Dissertation for the degree of Doctor of Science in Technology on 23 February 2001.



External examiners:
Pekka Kilpeläinen (University of Kuopio)
Heikki Kälviäinen (Lappeenranta University of Technology)
Opponent:
Pekka Orponen (University of Jyväskylä)

Abstract:
This thesis focuses on the analysis of irregular hierarchical visual objects. The main approach involves modelling the objects as unordered attribute trees. A tree presents the overall organization, or topology, of an object, while the vertex attributes describe further visual properties such as intensity, color, or size. Techniques for extracting, matching, comparing, and interpolating attribute trees are presented, principally aiming at systems that can learn to recognize objects. Analysis of weather radar images has been the pilot application for this study, but the main ideas are applicable in structural pattern recognition more generally.

The central original contribution of this thesis is the Self-Organizing Map of Attribute Trees (SOM-AT) which demonstrates how the proposed tree processing techniques - tree indexing, matching, distance functions, and mixtures - can be embedded in a learning system; the SOM-AT is a variant of the Self-Organizing Map (SOM), the neural network model invented by Prof. Teuvo Kohonen. The SOM is especially suited to visualizing distributions of objects, classifying objects and monitoring changes in objects. Hence, the SOM-AT can be applied in the respective tasks involving hierarchical objects. More generally, the proposed ideas are applicable in learning systems involving comparisons and interpolations of trees.

The suggested heuristic index-based tree matching scheme is another independent contribution. The basic idea of the heuristic is to divide trees to subtrees and match the subtrees recursively by means of topological indices. Given two attribute trees, the larger of which has N vertices, and the maximal child count (out-degree) is D vertices, the complexity of the scheme is only $O(N \log D)$ operations while exact matching schemes seem to have at least quadratic complexity: $O(N^{2.5})$ operations in checking isomorphisms and $O(N^3)$ operations in matching attribute trees. The proposed scheme is efficient also in terms of its "hit rate", the number of successfully matched vertices. In matching two random trees of $N \leq 10$ vertices, the number of heuristically matched vertices is on average 99 % of the optimum, and the accuracy for trees of $N \leq 500$ vertices is still above 90 %.

The feasibility of the proposed techniques is demonstrated by database querying, monitoring, and clustering of weather radar images. A related tracking scheme is outlined as well. In addition to weather radar images, a case study is presented on Northern light (*Aurora Borealis*) images. Due to the generic approach, there are also medical and geographical applications in view.

Auroral monitoring network: from all-sky camera system to automated image analysis

Mikko Syrjäsuo

Dissertation for the degree of Doctor of Science in Technology on 28 November 2001.

External examiners:
Ari Visa (Tampere University of Technology)
Jøran Moen (University of Oslo)
Opponents:
Jøran Moen (University of Oslo)
Jussi Parkkinen (University of Joensuu)

Abstract:
Auroras occur in the auroral zones encircling the magnetic poles of the Earth. Precipitating particles collide with atmospheric atoms and molecules in the ionosphere and produce light. In the northern hemisphere, we call this light aurora borealis or the Northern Lights. The solar wind, the magnetosphere and the ionosphere are electromagnetically coupled, which means that auroral activity is a direct consequence of the dynamic plasma processes that take place in the near-Earth space. Thus, the aurora provides us a way to monitor the Earth s plasma environment.

This thesis focuses on developing auroral imaging technology and data processing. The contribution to auroral instrumentation is a unique state-of-the-art network of all-sky camera stations that acquire images of the aurora in Fennoscandia and Svalbard. All stations operate autonomously and can also be remote controlled. The computer-controlled imagers utilise fish-eye lenses and image intensifiers to capture the whole sky into one digital image nominally at 20-s intervals during dark time. Also, calibration efforts have been initiated, which results in higher quality data than earlier.

The auroral data has traditionally been analysed manually, which is not only tedious but also prone to errors and is not easily repeatable. This thesis contributes the very first steps in combining contemporary image processing and machine vision techniques in analysing auroral images automatically. Techniques based on sample auroral images, shape skeletons and attribute trees are developed to classify the inherently noisy and ambiguous images. Time-series of images are analysed by utilising an auroral arc tracker.

The feasibility of the proposed techniques is demonstrated by comparing their performance to that of a human expert. The amount of required manual labour was vastly reduced by using the machine vision methods in searching for auroral arcs for a statistical study. Also, the auroral occurrence for one year was determined based on classifying 180000 images completely automatically.

Theses

Licentiate of Science in Technology

2000

Cristescu, Razvan
Applications of independent component analysis in telecommunications and image processing

Malaroiu, Simona
Applications of independent component analysis

Vesanto, Juha
Using SOM in data mining

2001

Hakulinen, Aimo
Riippulukkoleimojen automatisoitu visuaalinen laaduntarkastus (Automated visual quality control of padlock stamps)

Master of Science in Technology

2000

Aksela, Matti
Handwritten character recognition: A palm-top implementation and adaptive committee experiments

Engdahl, Kim
Avainhallinto usean korttien liikkeellelaskijan yleisöpuhelinmaksamisessa (Key management in public payphone payment systems with multiple card issuers)

Eriksson, Mika
Ortokuvien ja rasterikarttojen Internet jakelu (Internet distribution of orthoimages and raster maps)

Haanaho, Ismo
Frequency estimation of radiosonde signal

Inki, Mika
Topographic independent component analysis: Theory and applications

Koivulainen, Jani
Implementation of the type system for OTNSDL, an object-oriented SDL-variant

Laakso, Sami
Implementation of content-based WWW image search engine

Leppänen, Jaakko
A uniform web profiling system

Lybeck, Christer
Routing of partially protected routes in a transmission network management system

Mikkelä, Petri
Metsien kaukokartoitus ERSI SAR -tutkakuvilla (Remote sensing of forests using ERSI SAR -radarimages)

Mikkola, Topi
Algorithm library for large scale vehicle routing

Mäkinen, Tommi K.
Intelligent supplementary services for electronic information browsers

Nousiainen, Sami
Simulation methods for UMTS network

Ostrovskij, Jori
Software unit testing

Pantsar, Tuomas
Detection of surface cracks and pipe joints in digital sewer images

Parviainen, Jukka
Data mining for finding surface defects in steel strips

Pirskanen, Janne
Automaattisen viantunnistusjärjestelmän kehittäminen (Developing automatic recognition of hardware failures)

Pitkänen, Anna
Methods for improving spectrum efficiency in fixed radio networks

Puittinen, Mikael
Implementing project management tools in a world wide web based engineering data management system

Rinta-Koski, Olli-Pekka
Calculating margin requirements for derivative clearing accounts

Saarela, Antti
Paperin ja metallin pintavikakuvien analysointi (Analysis of surface defects in paper and metal)

Siponen, Markus
Automaattisia jälkitulkintamenetelmiä hierarkisen tietoaineiston tutkimiseen itseorganisoivan kartan avulla (Post processing tools for self-organising map and hierarcial data structures)

Siren, Juuso
Adaptiivinen menetelmä kuvalevystä luetun röntgenkuvan parantamiseen (Adaptive image enhancement method for CR images)

Syrjälahti, Mikko
Analyzing organizational emergence with self-organizing maps

Venna, Jarkko
Menetelmä moniulotteisen tietoaineiston kasaumarakenteiden havainnollistamiseksi värien avulla (A method for visualizing the cluster structure of multidimensional data using colours)

2001

Ahola, Jussi
Analysis of mechanical properties of hot rolled steel strips using SOM-based data mining

Alaluusua, Juha
Adaptive multi-rate wideband speech codec in 3G media gateway

Bäckström, Tom
Development of and analysis environment for clinical analysis of continuous speech

Girdziusas, Ramunas
Discriminative on-line recognition of isolated handwritten characters

Halinen, Jussi
Algorithm for the correction of ECG baseline

Honkela, Antti
Nonlinear switching state-space models

Korhonen, Ilkka
Porting operating system to mobile communication device: case study using EPOC operating system and ARM processor

Nirhola, Marko
Modeling of speech coding effects in an acoustic echo canceller of a 3G mobile network

Parviainen, Olli
Ennakoivaan simulointiin perustuva panostuotannon optimointi (Batch production optimization by evolutionary computation and predictive simulation)

Peltonen, Jaakko
Self-organizing maps in learning metrics

Raiko, Tapani
Hierarchical nonlinear factor analysis

Reunanen, Juha
Feature selection for classification of metal strip surface defects

Sundell, Jonne
Implementing tandem free operation over an IP based network

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. 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, but aside from this, part of the field has developed into 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. The computational methods have merged with other related fields like advanced statistics, pattern recognition, signal and data analysis, machine learning, and artificial intelligence.

In such a situation, it may be difficult to define what neurocomputing really means. The way we understand it is that we build *empirical data based models of complex systems*. Natural data has properties such as nonlinearity, nongaussianity, and dynamic 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 that although the models are not available in closed form, the intrinsic features 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 – the SOM method alone has been used in more than 4400 scientific works. 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

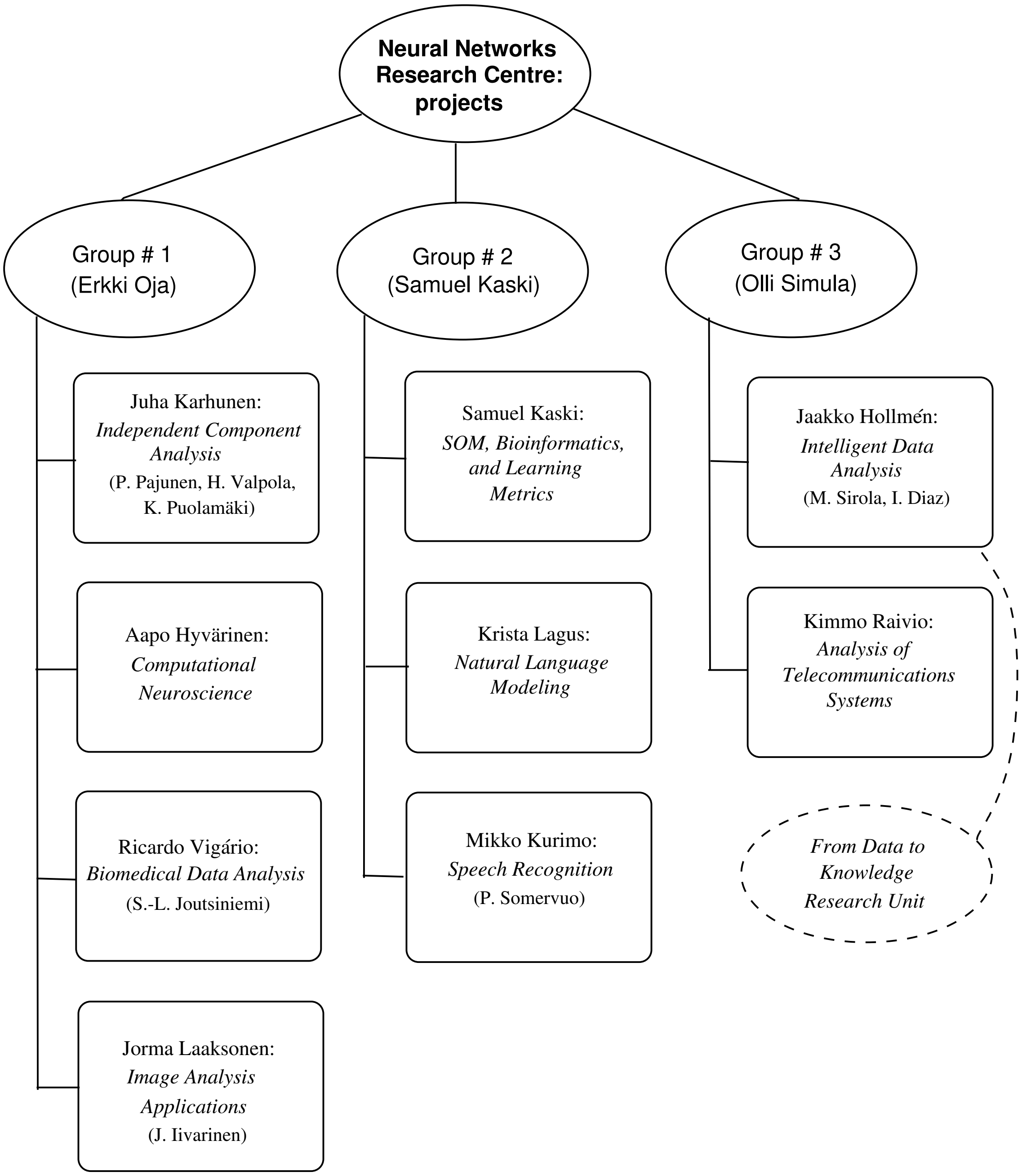


Figure 1.1: The Neural Networks Research Centre consists of three major groups, each having a number of smaller project groups. The leader of each group and the research topic are marked within each box, as well as the names of the post-doctoral researchers within each project. The dotted line indicates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

projects. Typically, these project groups consist of senior researchers, graduate students, and undergraduate students. 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. There are also some quite new research initiatives that do not fit under any of the projects;

at the end of this report, there is a separate Chapter for the other research done in our Centre.

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

Chapter 2

Independent component analysis and blind source separation

Erkki Oja, Juha Karhunen, Aapo Hyvärinen, Petteri Pajunen, Ricardo Vigário, Harri Valpola, Jaakko Särelä, Ella Bingham, Mika Inki, Antti Honkela, Tapani Raiko, Karthikesh Raju, Alexander Ilin, Rãzvan Cristescu, Simona Mãlãroiu, Kimmo Kiviluoto, Mika Ilmoniemi

2.1 Introduction

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.

The technique of ICA is a relatively new invention. It was for the first time introduced in early 1980's in the context of artificial neural networks. In mid-1990's, some highly successful new algorithms were introduced by several research groups, together with impressive demonstrations on problems like the cocktail-party effect, where the individual speech waveforms are found from their mixtures. ICA became one of the exciting new topics both in the field of neural networks, especially unsupervised learning, and more generally in advanced statistics and signal processing.

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. That earlier work has been reported in the previous Triennial reports of our laboratory from 1994 to 1999.

In the reporting period 2000 - 2001, our ICA research 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: "Bayesian ensenble learning of generative models" (a project led by prof. Juha Karhunen and Dr. Harri Valpola), "Biomedical data analysis" (a project led by Dr. Ricardo Vigario) and "Computational neuroscience" (a project led by Doc. Aapo Hyvärinen). There is also a separate Chapter for various applications of ICA.

This Chapter covers some theoretical projects that ended during the reporting period. Then, the EU project BLISS is reviewed, and the chapter ends with a description of the ICA2000 international workshop organized by our group.

2.2 Theoretical advances

Local ICA

Juha Karhunen, Simona Mãlãroiu, Mika Ilmoniemi

In standard Independent Component Analysis (ICA), a linear data model is used for a global description of the data. Even though linear ICA yields meaningful results in many cases, it can provide a crude approximation only for general nonlinear data distributions. We have proposed a new structure [1], where local ICA models are used in connection with a suitable grouping algorithm clustering the data. The clustering part is responsible for an overall coarse nonlinear representation of the data, while linear ICA models of each cluster are used for describing local features of the data. The goal is to represent the data better than in linear ICA while avoiding computational difficulties related with nonlinear ICA.

Several data grouping methods were considered, including standard K-means clustering, self-organizing maps, and neural gas. In the journal paper [1] summarizing the results of this research line, connections to existing methods are discussed, and experimental results are given for artificial data and natural images. Furthermore, a general theoretical framework encompassing a large number of methods for representing data is introduced in [1]. These range from global, dense representation methods to local, very sparse coding methods. The proposed local ICA methods lie between these two extremes.

The FastICA algorithm for complex valued signals

Ella Bingham and Aapo Hyvärinen

The purpose of this project was to extend the applicability of the FastICA algorithm [4, 5] to complex-valued signals.

Separation of complex valued signals is a frequently arising problem in signal processing: frequency-domain implementations involving complex valued signals have advantages over time-domain implementations. Especially in the separation of convolutive mixtures it is a common practice to Fourier transform the signals, which results in complex valued signals. It is assumed that the original, complex valued source signals are mutually statistically independent, and the problem is solved by the independent component analysis (ICA) model.

We presented a fast fixed-point type algorithm that is capable of separating complex valued, linearly mixed source signals. The computational efficiency of the algorithm was shown by simulations. Also, the local consistency of the estimator given by the algorithm was proved. The results are presented in detail in [2, 3].

Overlearning in ICA

Jaakko Särelä and Ricardo Vigário

All ICA algorithms face the classical overlearning problem, if the number of data points is too low compared to the dimension of the data. This overlearning problem is studied in [6]. In the extreme case, where there are only as many data points as there are dimensions, the ICA algorithms result in the extreme maximization of their cost functions, totally independent of the data. In case of higher-order ICA algorithms this results in signals that are almost zero everywhere except for a single spike. This overlearning problem is usually quite easy to overcome by collecting more data. Even more efficient way is the reduction of the data dimension by PCA, when the dimension of the data exceeds the number of underlying sources.

If the sources are not independent in time, the problem becomes more severe. This is because the additional information each new data point gives is reduced, thus diminishing the *effective* number of data points. This temporal dependence, with its associated low frequencies, privileges the appearence of *bumps* rather than the spikes. This behaviour is very common to magnetoencephalographic signals (see Chapter 6). The paper [6] presents several preprocessing techniques as well as modifications to the classical ICA algorithms to overcome the problem of bumps.

References

[1] J. Karhunen, S. Mãlãroiu, and M. Ilmoniemi, Local linear independent component analysis based on clustering. *Int. J. Neural Systems*, Vol. 10, No. 6, 2000, pages 439-451.

[2] Ella Bingham and Aapo Hyvärinen. A fast fixed point algorithm for Independent Component Analysis of complex valued signals. *International Journal of Neural Systems*, 10(1):1–8, February 2000.

[3] Aapo Hyvärinen, Juha Karhunen and Erkki Oja. *Independent Component Analysis*, Wiley Interscience, 2001.

[4] Aapo Hyvärinen and Erkki Oja. A fast fixed-point algorithm for Independent Component Analysis. *Neural Computation*, 9:1483–1492, 1997.

[5] Aapo Hyvärinen. Fast and robust fixed-point algorithms for Independent Component Analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634, May 1999.

[6] J. Särelä and R. Vigário. The problem of overlearning in high-order ICA approaches: analysis and solutions. Proceedings of the international workshop on artificial neural networks (IWANN'01), (Granada,Spain), pages 818–825, 2001.

2.3 The European joint project BLISS

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

Our laboratory is one of the five participants in a large European joint project on Blind Source Separation and Applications, abbreviated BLISS. The project covers the three year period between June 2000 and June 2003, and its total funding is 1.2 million euros. It belongs 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 are:

- 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).

In addition, the project has both industrial and scientific advisory board members.

The project divides into two major parts: Theory and Algorithms, and Applications. The first part Theory and Algorithms consists of three subprojects, which are Linear ICA, Nonlinear Separation, and Nonlinear BSS (Blind Source Separation). Our laboratory is involved in the last two subprojects. The second major part Applications has two subprojects, Biomedical Applications and Acoustic Mixtures, and our laboratory participates in the first subproject Biomedical Applications.

The kick-off meeting of the whole project was arranged by our laboratory in context with the ICA2000 workshop held in Espoo in June 2000. The project has official meetings twice a year, with a review meeting after each passed year. The first review meeting was held in Granada, Spain, in June 2001. The reviewers assessed there this project as an excellent one. The third meeting was held in San Diego in December 2001 in context with the ICA2001 workshop. There has also been some visiting researchers between the participating laboratories for deepening practical co-operation. A major event will be the European Meeting on Independent Component Analysis to be held in Vietri sul Mare, Italy, in February 2002. It is largely organized by the participants of the BLISS project, in particular Prof. Erkki Oja, together with Italian researchers. This meeting also serves as a winter school for graduate students.

The research results achieved by our laboratory using the BLISS project funding are described in the chapters "Bayesian ensemble learning for 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 confidential deliverables and reports of the project. More information on the BLISS project is available on its homepage [1].

References

[1] Homepage of the BLISS project `http://www.bliss-project.org/`.

2.4 New book: Independent Component Analysis, by A. Hyvärinen, J. Karhunen, and E. Oja (Wiley, 2001)

The original articles on ICA have been published during the past 20 years in a large number of journals and conference proceedings in the fields of statistics, signal processing, artificial neural networks, information theory, and various application fields. Some edited collections of articles as well as some monographs on ICA, blind source separation, and related subjects have appeared recently. However, while highly useful for their intended readership, these existing texts typically concentrate on some selected aspects of the ICA methods only. In the brief scientific papers and book chapters, mathematical and statistical preliminaries are usually not included, which makes it very hard for a wider audience to gain full understanding of this fairly technical topic. A comprehensive and detailed text-book was missing, which would cover both the mathematical background and principles, algorithmic solutions, and practical applications of the present state-of-the-art of ICA.

The three authors (Dr. Aapo Hyvärinen and professors Juha Karhunen and Erkki Oja) took up the task of writing such a text-book. The ordering of the names reflects the number of chapters that each author wrote, with Aapo Hyvärinen doing the bulk of the work. All the text was carefully read by all the authors, with many rewritings, resulting in the finished manuscript. The project lasted from summer 1999 to autumn 2000, and the book "Independent Component Analysis" [1] came out in spring 2001, published by Wiley-Interscience in the Wiley Series on Adaptive and Learning Systems for Signal Processing, Communications, and Control. The series editor, prof. Simon Haykin, was quite influential in convincing us to write the book and in seeing it through the various stages of publication.

This book provides a comprehensive introduction to ICA as a statistical technique. The emphasis is on the fundamental mathematical principles and basic algorithms. Much of the material is based on the original research conducted in the authors' own research group, which is naturally reflected in the weighting of the different topics. We give a wide coverage especially to those algorithms that are scalable to large problems, having a large number of mixtures and sources available. These will be increasingly used in near future, when ICA moves from the toy problems used in theoretical research into practical real world problems. Respectively, somewhat less emphasis is given to signal processing methods involving convolutive mixtures, delays, and other blind source separation techniques than ICA.

This book serves as a fundamental introduction to ICA. It is expected that the readership will be from a variety of disciplines such as statistics, signal processing, neural networks, cognitive science, information theory, artificial intelligence, mathematics, and engineering. The book is suitable for a graduate level university course on ICA, which is facilitated by the exercise problems and computer assignments given throughout the book, as well as by the lecture slides that are now available, resulting from the course given by J. Karhunen and E. Oja at HUT in autumn 2001.

For more details of the book, see `http://www.cis.hut.fi/projects/ica/book/`.

References

[1] Aapo Hyvärinen, Juha Karhunen and Erkki Oja. *Independent Component Analysis*, Wiley Interscience, 2001.

2.5 ICA2000: the Second International Workshop on Independent Component Analysis and Blind Source Separation

Erkki Oja, Juha Karhunen, Visa Koivunen, Petteri Pajunen, Aapo Hyvärinen, Jukka Iivarinen

Beginning from mid-nineties, Independent Component Analysis (ICA) has become a very popular research topic in the intersection of signal processing and neural networks. The first major scientific meeting exclusively devoted to ICA was the International Workshop on Independent Component Analysis and Blind Source Separation (ICA99), organized in January 1999 in Aussois in the French Alps. The workshop was highly successful, with every major ICA research group represented.

During that meeting it became clear that a continuation of the workshop was needed, and the Finnish group at Helsinki University of Technology undertook the job of organizing the second workshop on ICA and BSS. As for the timing, the general opinion was that an annual meeting may be too frequent but a two year interval is too long. We settled for a compromise of one and a half year interval, also motivated by the fact that June is definitely preferable over January for a conference held in northern Europe.

The Second International Workshop on Independent Component Analysis and Blind Source Separation (ICA2000) was held on June 19 - 22, 2000, on the beautiful Hanasaari island in Espoo. It was a major effort for our ICA group. The General Chairman was Prof. Erkki Oja, the Program Chair Prof. Juha Karhunen, the Local Arrangements Chair Prof. Visa Koivunen, the Publications Chair Prof. Petteri Pajunen, the Publicity Chair Dr. Aapo Hyvärinen, and the Financial Chair Dr. Jukka Iivarinen. Mr. Jaakko Särelä and Mr. Markus Peura were in charge of the graphical design and homepages of the workshop, and a large number of the members of our local ICA research community acted as volunteer assistants in the various conference tasks. The core task, selecting the papers and making up the program, was expertly handled by the International Program Committee, consisting of 26 eminent ICA researchers from Finland, France, Germany, UK, Portugal, USA, Canada, Japan, and P.R.China.

The workshop program contained four invited papers by eminent scientists working on ICA, BSS, or closely related fields, as well as 99 high-quality technical papers. The invited talks were "Source separation: from dusk till dawn" by prof. C. Jutten (Institut National Polytechnique de Grenoble, France), "Multiuser detection: an overview and a new result" by prof. U. Madhow (University of Santa Barbara, USA), "Nature vs. math: interpreting ICA in light of recent work in harmonic analysis" by prof. D. Donoho (Stanford University, USA), and "Independent component analysis of biomedical signals" by prof. T. Sejnowski (Salk Institute, San Diego, USA). The regular sessions ranged from theoretical ones to several applications. The number of registered participants was about 120. This was a well focussed, high level workshop covering all the main aspects of ICA and blind source separation.

Full contents of the papers are included in the Proceedings [1], now also available in CD Rom. The Web page `www.cis.hut.fi/ica2000/` gives some more information, including several photos of the main events.

The two ICA meetings had a succession in December, 2001, when the Third International Workshop on ICA and BSS was held in San Diego, USA. The next ICA workshop, to be held in Nara, Japan, is already at the planning stage.

Figure 2.1: Participants of the ICA workshop on the lawn of Hanasaari Cultural Centre

References

[1] *Proceedings of the Second International Workshop on Independent Component Analysis and Blind Signal Separation*. P. Pajunen and J. Karhunen, Editors. Helsinki University of Technology, Neural Networks Research Centre, 2000.

Chapter 3

Applications of independent component analysis

Erkki Oja, Juha Karhunen, Aapo Hyvärinen, Petteri Pajunen, Ricardo Vigário, Harri Valpola, Jaakko Särelä, Ella Bingham, Mika Inki, Antti Honkela, Tapani Raiko, Karthikesh Raju, Alexander Ilin, Rãzvan Cristescu, Simona Mãlãroiu, Kimmo Kiviluoto, Mika Ilmoniemi, Mark Girolami, Ata Kabán, Maria Funaro

3.1 Decision trees using independent component analysis

Petteri Pajunen and Mark Girolami

Decision trees are typically used as computationally efficient representations in classification or regression. For example binary trees allows one to reach a leaf node in $\log_2 n$ decisions when the tree contains n nodes in total. If the decisions are easy to compute, this may result in a very light computational process.

Typical problems involving decision trees deal with multivariate data where the components are usually attribute values with a clear interpretation. Decisions can be simply implemented as threshold values on one of the attributes. An example is an attribute $height(k)$ which is the height of person k. A decision could be a threshold $height(k) > 175cm$, which would split the data in two classes based on height.

One of the key properties of Independent Component Analysis is the ability to find linear combinations of multivariate data that have certain information-theoretic properties. Often these linear combinations represent underlying sources, which cannot be directly observed. If the data contains such hidden sources, decision based on thresholding observed data components may not be very effective.

In [1], single-component ICA was used to implement binary decisions in a decision tree. The key idea is that the independent components have more structure than the observed components, and therefore can be expected to be better candidates for linear threshold decisions. Since only single ICA component needs to be computed at each tree node, computational complexity resulting from a large number of ICA components can be avoided.

References

[1] P. Pajunen and M. Girolami. Implementing Decisions in Binary Decision Trees using Independent Component Analysis. In *Proceedings of the 2nd Int. Workshop on Independent Component Analysis and Blind Source Separation*, pages 483–487, Espoo, Finland, June 2000.

3.2 ICA for text mining

Ella Bingham, Ata Kabán, Mark Girolami

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.

As a preprocessing step, the text stream is split into short windows, and from each window a T-dimensional vector is formed, where T is the size of the vocabulary; T is typically several thousands of terms. The i-th element of the vector indicates (some function of) the frequency of the i-th vocabulary term in the window. The high dimensionality of the data is reduced by singular value decomposition, as is often done before applying ICA-type algorithms on the data.

Our method of finding the topics of dynamical text is based on the *complexity pursuit* algorithm presented by Hyvärinen [3]. This algorithm is a generalization of projection pursuit to time series, and it is closely related to ICA. The algorithm gives us the directions into which the multidimensional time series data should be projected; these projections are then our estimates of the topic time series.

We give here an example of analyzing chat line discussions[1]. Results on newsgroup data with comparisons to other algorithms are presented in [1] and [2]. Figure 3.1 shows how different topic time series are activated at different times. We can see that the topics clearly are autocorrelated in time. The time span of Figure 3.1 is almost 24 hours; some topics are more or less persistent during the whole period and some will come up again after a few hours.

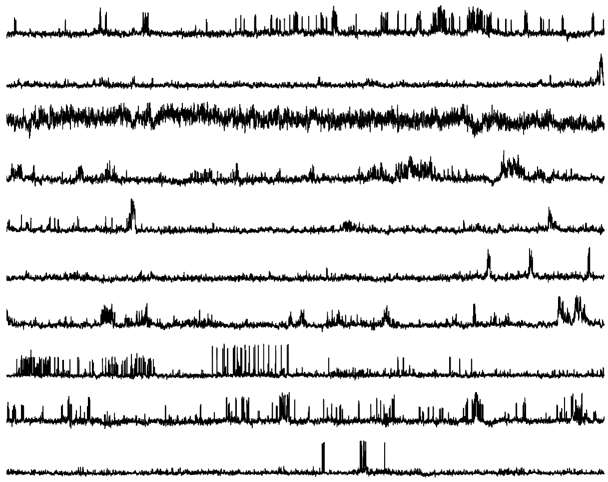


Figure 3.1: Activity of topics (vertical axis) in each chat window (horizontal axis). The uppermost time series corresponds to topic 1, the second to topic 2 etc.

[1] `http://www.cnn.com/chat/channel/cnn_newsroom`

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
jackson	site	bush	religion	violenc
sharpton	web	ashcroft	god	report
child	net	vote	jesu	youth
stori	word	kennedi	bibl	children
drudg	parent	presid	religi	gun
rainbow	nanni	cnn	life	point
monei	internet	time	follow	home
mistress	block	gore	read	drug
coalition	kid	question	stori	famili
tonight	system	elect	univers	satcher
pregnant	access	god	exist	health
affair	child	senat	faith	risk
black	base	power	man	factor
chenei	chat	thing	book	surgeon
jessi	page	fact	earth	prevent
Topic 6	**Topic 7**	**Topic 8**	**Topic 9**	**Topic 10**
flag	california	join	tax	free
move	power	discuss	cut	liber
citi	electr	est	exempt	opinion
ncaa	energi	tonight	monei	religion
offici	blackout	room	gop	form
atlanta	state	studio	hous	polit
count	deregul	cnn	congress	conserv
game	compani	conserv	pay	birth
night	crisi	american	interest	philosophi
georgia	price	nea	recess	establish
chang	plant	union	payer	narrow
lose	util	keen	secur	restrict
confeder	order	type	henri	independ
hehe	home	chat	hypocrit	orthodox
chenei	cost	newsroom	hyde	bound

Table 3.1: Keywords of chat line discussion topics related to the time series in Figure 3.1.

The most significant terms of the topics are listed in Table 3.1. It is seen that each keyword list indeed characterizes one distinct topic quite clearly. Topic 1 deals with Jesse Jackson and his illegitimate child, topic 2 is about parental control over children's web usage and topic 3 is a general discussion about G.W. Bush. Topic 4 is a religious discussion, topic 5 deals with problems of the youth such as violence and drug abuse, and topic 6 is about the controversial flag of the state of Georgia, US, due to which the NCAA basketball games risked cancellation in Atlanta. Topic 7 involves the energy shortage in California, topic 8 corresponds to comments given by the chat line moderator, topic 9 is about taxation and topic 10 is a short discussion dealing with the values of the politicians in the US.

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] Bingham, E. Topic identification in dynamical text by extracting minimum complexity time components. In: *Proc. 3rd International Conference on Independent Component Analysis and Blind Signal Separation (ICA2001)*, December 9–13, 2001, San Diego, CA, USA, pp. 546–551.

[2] Bingham, E., A. Kabán, and M. Girolami. Topic identification in dynamical text by complexity pursuit. *Neural Processing Letters*. Submitted.

[3] Hyvärinen, A. Complexity pursuit: separating interesting components from time-series. *Neural Computation* **13**(4), 883–898.

3.3 ICA for analyzing financial time series

Simona Mãlãroiu, Kimmo Kiviluoto, Erkki Oja

It is a tempting alternative to try ICA on financial data. There are many situations in which parallel financial time series are available, such as currency exchange rates or daily returns of stocks, that may have some common underlying factors. ICA might reveal some driving mechanisms that otherwise remain hidden. If one could find the maximally independent mixtures of a set of original stocks, i.e. independent portfolios, this might help in minimizing the risk in the investment strategy.

Another promising application is the prediction of financial time series. The ICA transformation tends to produce component signals $s_j(t)$ that can be compressed with fewer bits than the original signals $x_i(t)$. They are thus more structured and regular. This gives motivation to try to predict the signals $x_i(t)$ by first going to the ICA space, doing the prediction there, and then transforming back to the original time series. The prediction can be done separately and with a different method for each independent component, depending on its time structure. Hence, some interaction from the user may be needed in the overall prediction procedure.

An algorithm for this was suggested in [1, 2, 3]. After subtracting the mean of each time series and prewhitening, the independent components $s_j(t)$ and the mixing matrix $\mathbf{A}$ are estimated using the FastICA algorithm. For each component $s_j(t)$, a suitable nonlinear filtering is applied to reduce the effects of noise. Each smoothed independent component is then predicted separately for a number of steps into the future, for instance using some method of AR modeling. The final predictions for the original time series $x_i(t)$ are obtained by inverting the ICA transformation.

To test the method, we applied our algorithm on a set of 10 foreign exchange rate time series. The results were promising, as the ICA prediction performed better than direct prediction. Fig. 3.2 shows an example of prediction using our method. The upper figure represents one of the original time series (mixtures) and the lower one the forecast obtained using ICA prediction for a future interval of 50 time steps. The algorithm seemed to predict very well especially the turning points.

References

[1] Mãlãroiu, S., Kiviluoto, K. and Oja, E.: Time series prediction with Independent Component Analysis. *Proc. Int. Conf. on Advanced Investment Technology*, December 19 - 21, 1999, Queensland, Australia (1999).

[2] Mãlãroiu, S., Kiviluoto, K. and Oja, E.: ICA preprocessing for time series prediction. *Proc. 2nd Int. Workshop on Indep. Comp. Anal. and Blind Source Separation*, June 19 - 22, 2000, Helsinki, Finland, pp. 453 - 457 (2000)

[3] Oja, E., Kiviluoto, K. and Mãlãroiu, S.: Independent component analysis for financial time series. *Proc. IEEE 2000 Symp. on Adapt. Systems for Signal Proc., Comm. and Control AS-SPCC*, Oct. 1 - 4, 2000, Lake Louise, Canada, pp. 111 - 116 (2000).

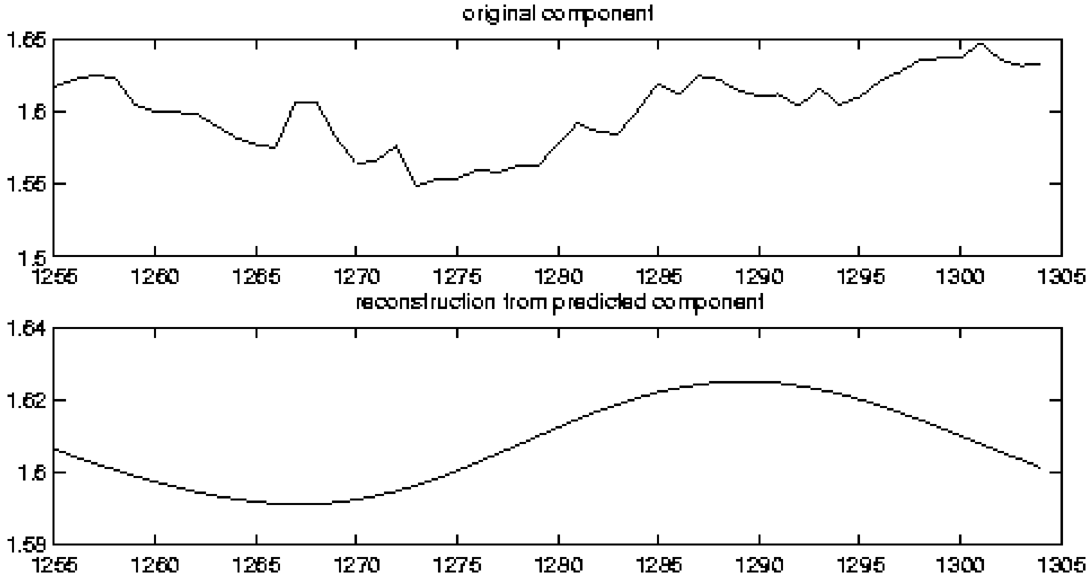


Figure 3.2: Prediction of real-world financial data: the upper figure represents the actual future outcome of one of the original mixtures and the lower one the forecast obtained using ICA prediction for an interval of 50 values.

3.4 ICA for astronomical data

Maria Funaro, Erkki Oja, Harri Valpola

A new approach in our group is to try ICA for removing artifacts from astronomical telescope images. This problem was introduced to us by Ms. Maria Funaro who finalized her Ph.D. Thesis in our lab in spring 2001.

In modern astrophysics, one of the main research directions is understanding the dark matter in the universe. Of special interest are compact objects with substellar mass, such as black holes, dwarf stars, or planets. 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 (see [1]), the pixel luminosity variations over time are monitored, and using these time series the lensing events can yet be detected even in the case of unresolved stars.

A problem in the analysis of the images and luminosity variations is the presence of artefacts. One of the possible artefacts are the resolved or 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 possible physical events is one of the necessary steps in the analysis of pixel lensing data.

The new idea proposed by us [2] is to use ICA for the artefact detection and removal. This is motivated by the fact that for astrophysical data, the independence of the artefacts is often theoretically guaranteed, and also the linear mixing model holds exactly. This is an almost ideal application for ICA.

In the astrophysical data, we have a number of digital images, recorded over consequent nights when the conditions are favourable, and carefully calibrated for geometrical and photometric alignments. ICA is then used for this image set to reveal independent components that might be artefacts. An example is given in Fig. 3.3.

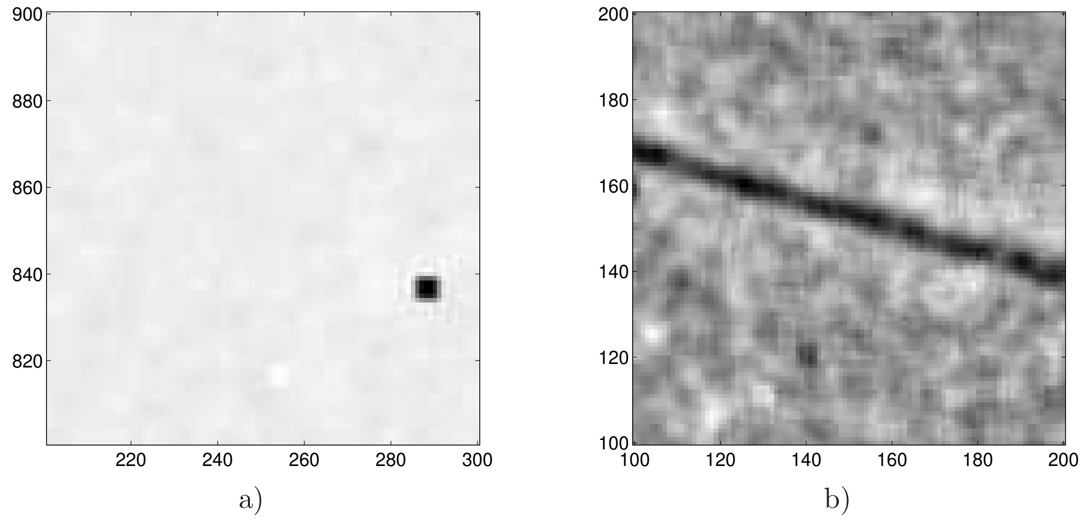


Figure 3.3: Two artefacts: (a) cosmic ray, (b) atmospheric effect.

References

[1] A. Gould, The theory of pixel lensing, *Astrophys. J.* 470, 1996, pp. 201 - 210.

[2] Funaro, M., Oja, E. and Valpola, H.: Artefact detection in astrophysical image data using Independent Component Analysis. *Proc. 3rd Int. Workshop on Indep. Comp. Anal. and Blind Source Separation*, Dec. 9 - 12, 2001, San Diego, USA, pp. 43 - 48 (2001).

3.5 ICA in CDMA communications

Rãzvan Cristescu, Karthikesh Raju, 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. These two requirements have given rise to a number of multiple access schemes [2]. The traditional ones are FDMA and TDMA schemes, based on the use of either non-overlapping frequency or time slots assigned to each user, respectively.

During the last years, various systems based on CDMA (Code Division Multiple Access) techniques [1, 2] have become popular. In CDMA, there is no disjoint division in frequency or time slots, but the users are identified by their unique codes. CDMA systems offer several advantages over FDMA and TDMA techniques. 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]. A serious problem is multipath propagation which is illustrated in Figure 3.4. In practice, one must typically resort to suboptimal techniques in processing CDMA signals for several reasons. Optimal methods may require prior information which is not available, they are usually computationally too demanding, and the assumptions made may not hold because of rapidly varying environment.

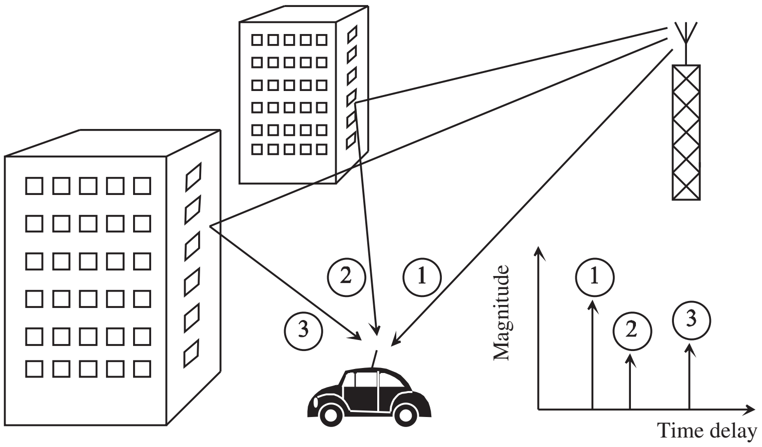


Figure 3.4: An example of multipath propagation in urban environment.

It turns out that direct sequence CDMA data can be cast in the form of a linear ICA/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. On the other hand, multipath propagation, possibly fading channels, and time delays make separation of the desired user's symbols a very challenging estimation problem which is more complicated than the standard linear ICA problem.

In the first stage of this project, we applied ICA and BSS methods to various problems in multiuser detection [1, 2], trying to take into account the available prior information whenever possible. In particular, we have considered estimation of the desired user's time delays in [5], estimation of fading channels in [4], and detection of the desired user's symbol sequence in [6]. The results are very promising, showing that ICA based methods can yield considerably better performances than more conventional methods based on second-order statistics. The work done during this stage is reviewed with the necessary background in Chapter 23 of the book [3].

In the second stage of the project, we have applied independent component analysis to blind suppression of interference caused by bit-pulsed jamming in a direct sequence CDMA communication system. This jamming problem is important 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 are reported in more detail in forthcoming conference papers [7, 8].

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] R. Cristescu, J. Joutsensalo, J. Karhunen, and E. Oja, A complexity minimization approach for estimating fading Gaussian channel in CDMA communications. In *Proc. of the 2nd Int. Workshop on Independent Component Analysis and Blind Source Separation (ICA2000)*, Espoo, Finland, June 19-22, 2000, pages 527-532.

[5] R. Cristescu, T. Ristaniemi, J. Joutsensalo, and J. Karhunen, CDMA delay estimation using a fast ICA algorithm. In *Proc. of the IEEE Int. Symp. on Personal, Indoor, and Mobile Communications (PIMRC'00)*, London, United Kingdom, September 17-19, 2000.

[6] R. Cristescu, T. Ristaniemi, J. Joutsensalo, and J. Karhunen, Blind separation of convolved mixtures for CDMA systems. In *Proc. of the X European Signal Processing Conference (EUSIPCO 2000)*, Tampere, Finland, September 5-8, 2000, pages 619-622.

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

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

Chapter 4

Bayesian ensemble learning of generative models

Harri Valpola, Antti Honkela, Juha Karhunen, Tapani Raiko, Xavier Giannakopoulos, Alexander Ilin, Erkki Oja

4.1 Bayesian modeling and ensemble learning

In unsupervised learning, the goal is to build a model which captures the statistical regularities in the observed data and provides a compact and meaningful representation for the data. From such a representation it is often much easier to understand the basic characteristics of the data than directly from the raw data.

Unsupervised learning methods are often based on a generative approach where the goal is to find a specific 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 good nonlinear model is a challenging task, because it is typically computationally much more demanding than for linear models, and the choice of an appropriate specific model is often difficult.

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

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

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

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

The parameters $\boldsymbol{\theta}$ of a particular model $\mathcal{H}_i$ are often estimated by seeking the peak value of a probability distribution. The non-Bayesian maximum likelihood (ML) method uses to this end the distribution $p(X|\boldsymbol{\theta},\mathcal{H})$ of the data, and the Bayesian maximum a posteriori (MAP) method finds the parameter values that maximize the posterior distribution $p(\boldsymbol{\theta}|X,\mathcal{H})$. 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 perform estimation by averaging over the posterior distribution $p(\boldsymbol{\theta}|X,\mathcal{H})$. This means that the estimates will be sensitive to regions where the probability mass is large instead of being sensitive to high values of the pdf [4]. One can even make use of the complete set of models by computing the predicted values as averages over the predictions given by several models, weighted by their respective probabilities. Such a procedure appropriately solves the issues related to the model complexity and choice of a specific model $\mathcal{H}_i$ among several candidates. In practice, however, the differences between the probabilities of candidate models are often very large, and hence it is sufficient to select the most probable model and use the estimates or predictions given by it.

A problem with fully Bayesian estimation is that the posterior distribution (4.1) has a highly complicated form except for in the simplest problems. Therefore it is too difficult to handle exactly, and some approximative method must be used. Ensemble learning, also called variational Bayes, is a recently developed [2, 3] approximative fully Bayesian method, which has become popular because of its good properties. It uses the Kullback-Leibler (KL) information between two probability distributions $q(v)$ and $p(v)$. The KL information is defined by the cost function

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

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

A key idea in ensemble learning is to minimize the misfit between the actual posterior pdf and its parametric approximation using the KL information. 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 minus 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 has been discussed from information theoretic point of view in [2].

In the following subsections, we first consider the use of ensemble learning in nonlinear principal component (factor) analysis as well as in nonlinear independent component analysis. The static model used in nonlinear factor analysis is then extended to nonlinear dynamic state-space models. In these problems, we have employed multilayer perceptron networks as a flexible family of models. We then discuss construction of hierarchic models from a set of basic blocks in context with ensemble learning, and present applications of the developed Bayesian methods to inferring missing values from data, to detection of changes in process states, and to analysis of biomedical MEG data.

References

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

[2] 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.

[3] 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.

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

4.2 Nonlinear factor analysis and independent component analysis

The linear principal and independent component analysis (PCA and ICA) model the data as having been generated by independent sources through a linear mapping. The difference between the two is that PCA restricts the distribution of the sources to be Gaussian, whereas ICA does not, in general, restrict the distribution of the sources.

We have applied ensemble learning to nonlinear counterparts of PCA and ICA where the generative mapping form sources to data is not restricted to be linear [1, 2, 3]. The general form of the model is

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

This can be viewed as a model about how the observations were generated from the sources. The vectors $\mathbf{x}(t)$ are observations at time t, $\mathbf{s}(t)$ are the sources and $\mathbf{n}(t)$ the noise. The function $\mathbf{f}(\cdot)$ is a parametrised mapping from source space to observation space. We have used multi-layer perceptron (MLP) network with tanh-nonlinearities to model the mapping $\mathbf{f}$:

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

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.

Dimension reduction by nonlinear factor analysis

Just as their linear counterparts, the nonlinear versions of PCA and ICA can be used for instance in dimension reduction and feature extraction. The difference between linear and nonlinear PCA is depicted in Figure 4.1. In the linear PCA the data is described with a linear coordinate system whereas in the nonlinear PCA the coordinate system is nonlinear. The nonlinear PCA and ICA can be used for similar tasks as their linear counterparts, but they can be expected to capture the structure of the data better if the data points lie in a nonlinear manifold instead of a linear subspace.

Model order selection is an important issue is real applications: how many sources are there, how complex nonlinearities should be used, etc. One benefit of ensemble learning is that the cost function can be used for optimizing model structure by simply minimizing the cost. A sample case is shown in Figure 4.2. The data was created by nonlinearly mixing five Gaussian random sources. Ensemble learning is actually able to shut down unused parts of the model. The weights of the MLP network corresponding to the extra sources converge to zero which means that the extra sources are effectively pruned away from the model. This explains why the cost function saturates after five sources.

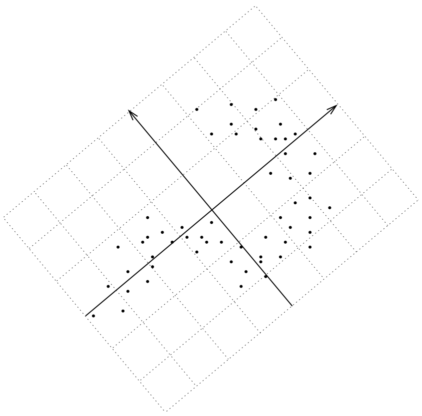


Figure 4.1: On the left hand side the data is described with a linear coordinate system. On the right hand side the coordinate system is nonlinear

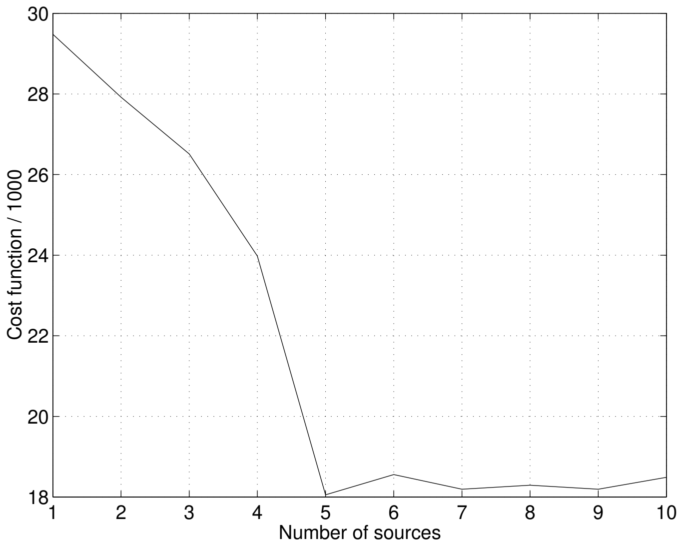


Figure 4.2: The value of the cost function is shown as a function of the number of sources. The MLP network had 30 hidden neurons. Ten different initialisations were tested to find the minimum value for each number of sources. The cost function saturates after five sources and the deviations are due to different random initialisation of the network

Nonlinear factor analysis is able to capture nonlinear structure in real data sets as can be seen in Figure 4.3. This data set consists of 2480 samples from 30 time series measured by different sensors from an industrial pulp process. An expert has preprocessed the signals by roughly compensating for time lags of the process which originate from the finite speed of pulp flow through the process. It appears that the data is quite nonlinear since the nonlinear factor analysis is able to explain as much data with 10 components as the linear factor analysis with 21 components.

Nonlinear independent factor analysis

Like in the linear case, Gaussian model for the sources results in rotational indeterminacy of the source space, whereas a non-Gaussian model for the sources fixes the rotation and

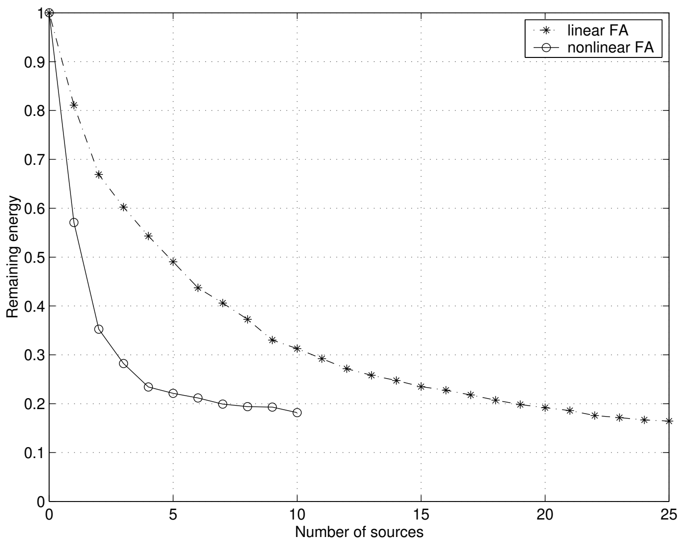


Figure 4.3: The graph shows the remaining energy in the process data as a function of the number of extracted components in linear and nonlinear factor analysis

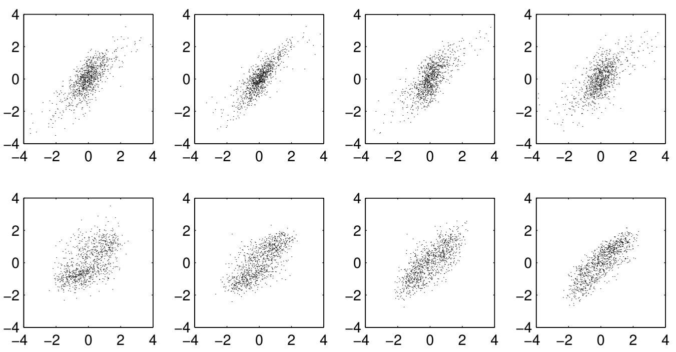


Figure 4.4: Original sources are on the x-axis of each scatter plot and the sources estimated by a linear ICA are on the y-axis. Signal to noise ratio is 0.7 dB

makes it possible to identify independent sources.

We have modeled the distribution of the sources by a mixture of Gaussian distributions. This density model is well suited for generative models and moreover, given enough Gaussians in the mixture, any density can be modeled by arbitrary accuracy using it.

Figures 4.4 and 4.5 illustrate the difference between linear and nonlinear ICA in the case where the mixing is nonlinear. The data set consisted of 1000 20-dimensional vectors which were created by nonlinearly mixing eight non-Gaussian independent random sources. Nonlinear model is clearly required in order to capture to underlying nonlinear manifold.

References

[1] 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.

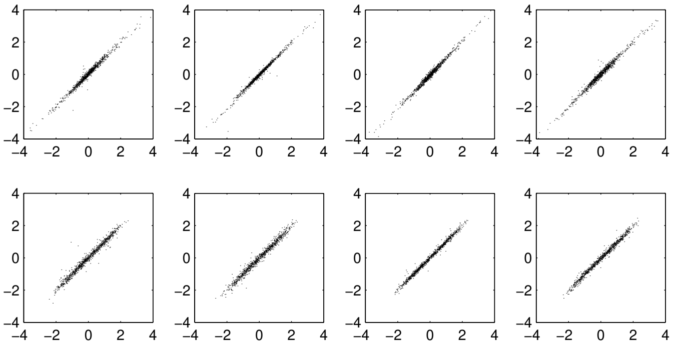


Figure 4.5: Nonlinear independent factor analysis is able to retrieve the original signals with small error. Signal to noise ratio is 17.3 dB

[2] Harri Valpola. Nonlinear independent component analysis using ensemble learning: Theory. In Petteri Pajunen and Juha Karhunen, editors, *Proceedings of the Second International Workshop on Independent Component Analysis and Blind Signal Separation, ICA 2000*, pages 251–256, Helsinki, Finland, June 2000.

[3] Harri Valpola, Xavier Giannakopoulos, Antti Honkela, and Juha Karhunen. Nonlinear independent component analysis using ensemble learning: Experiments and discussion. In Petteri Pajunen and Juha Karhunen, editors, *Proceedings of the Second International Workshop on Independent Component Analysis and Blind Signal Separation, ICA 2000*, pages 351–356, Helsinki, Finland, June 2000.

4.3 Nonlinear dynamic 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 $\mathbf{s}(t)$ [1, 2, 3]. This results in a state-space model where the sources can be interpreted as the internal state of the underlying generative process.

The model consists of the following set of equations:

$$\mathbf{x}(t) = \mathbf{f}(\mathbf{s}(t)) + \mathbf{n}(t) \tag{4.5}$$

$$\mathbf{s}(t) = \mathbf{g}(\mathbf{s}(t-1)) + \mathbf{m}(t)\,, \tag{4.6}$$

where $\mathbf{x}(t)$ are the observations, $\mathbf{s}(t)$ are the sources (states), $\mathbf{n}(t)$ and $\mathbf{m}$ are Gaussian observation and process noise and $\mathbf{f}(\cdot)$ and $\mathbf{g}(\cdot)$ are the nonlinear functions modeling the observations and dynamics, respectively.

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

$$\mathbf{g}(\mathbf{s}(\mathbf{t}-\mathbf{1})) = \mathbf{s}(t-1) + \mathbf{D}\tanh[\mathbf{C}\mathbf{s}(t-1) + \mathbf{c}] + \mathbf{d}\,. \tag{4.7}$$

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 over-fitting problems which have been major obstacles in developing this kind of unsupervised methods thus far. Potential applications for our method include prediction and process monitoring, control and identification. A process monitoring application is discussed in Section 4.5 in more detail.

References

[1] H. Valpola. Unsupervised learning of nonlinear dynamic state-space models. Technical Report A59, Lab of Computer and Information Science, Helsinki University of Technology, Finland, 2000.

[2] H. Valpola. *Bayesian Ensemble Learning for Nonlinear Factor Analysis*. PhD thesis, Helsinki University of Technology, Espoo, Finland, 2000. Published in Acta Polytechnica Scandinavica, Mathematics and Computing Series No. 108.

[3] H. Valpola and J. Karhunen. An unsupervised ensemble learning method for nonlinear dynamic state-space models. *Neural Computation*, 2002. Submitted, under revision.

4.4 Building blocks for ensemble learning

The methods for unsupervised learning of nonlinear latent variable models by ensemble learning introduced in Sections 4.2 and 4.3 represent a significant improvement in existing techniques in their ability to learn high-dimensional nonlinear representations of the data. However, the ability to scale these methods is limited by computational complexity which is roughly proportional to the number of data samples and the complexity of multiplying the matrices $\mathbf{A}$ and $\mathbf{B}$ in (4.4).

In order to be able to work with very large scale models, the computational complexity should be linear in the number of data samples and connections in the model. This can be achieved by utilizing the methods introduced in [1]. The idea is to construct large models from standardized building blocks which can be connected rather freely and can be learned by ensemble learning with local learning rules. Each block only needs to communicate with its neighbors which results in linear computational complexity.

The building blocks include continuous and discrete variables, summation, addition, nonlinearity and switching. Ensemble learning provides a cost function which can be used for updating the variables as well as optimising the model structure. The blocks are designed to fit together and to yield efficient update rules.

We have started to design a C++ library which can utilize these building blocks. The derivation of the cost function and learning rules is largely automatic which means that the user only needs to define the connections between the blocks. This research will be published in forthcoming papers.

References

[1] 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.

4.5 Applications

In this section, applications of nonlinear factor analysis (Section 4.2) and nonlinear state-space models (Section 4.3) are discussed. Application of the nonlinear state-space model to biomedical magnetoencephalographic data is discussed in Section 6.3.

Missing values

Generative models can usually easilly 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 as illustrated in Figure 4.6. 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.

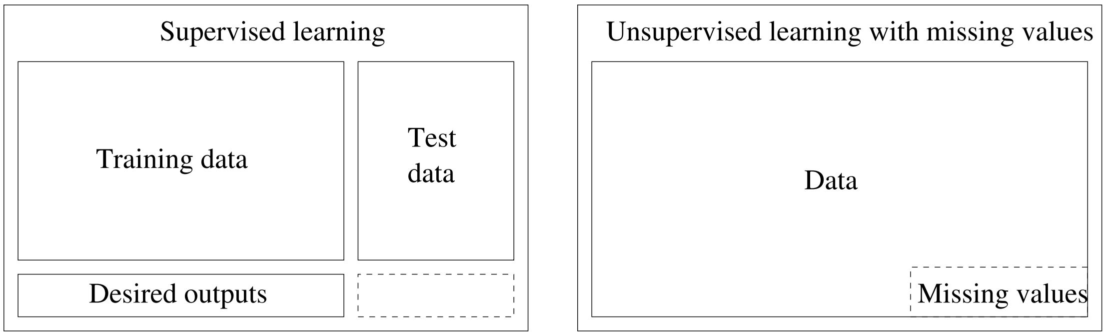


Figure 4.6: Unsupervised learning can be used for supervised learning by considering the outputs of the test data as missing values.

The quality of the reconstructions provides insight to the properties of different unsupervised models. The ability of self-organizing maps, linear principal component analysis and nonlinear factor analysis to reconstruct the missing values of various data sets have been studied in [2].

Figure 4.7 shows the results of reconstructing missing values in speech spectra. On the left, the reconstructed spectra are shown for a case where data were missing in patches and the data is not permutated. This corresponds to the upper row of the plots on the right. The other rows on the right show results for cases where the data was permutated and/or data was missing randomly instead of in patches. 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

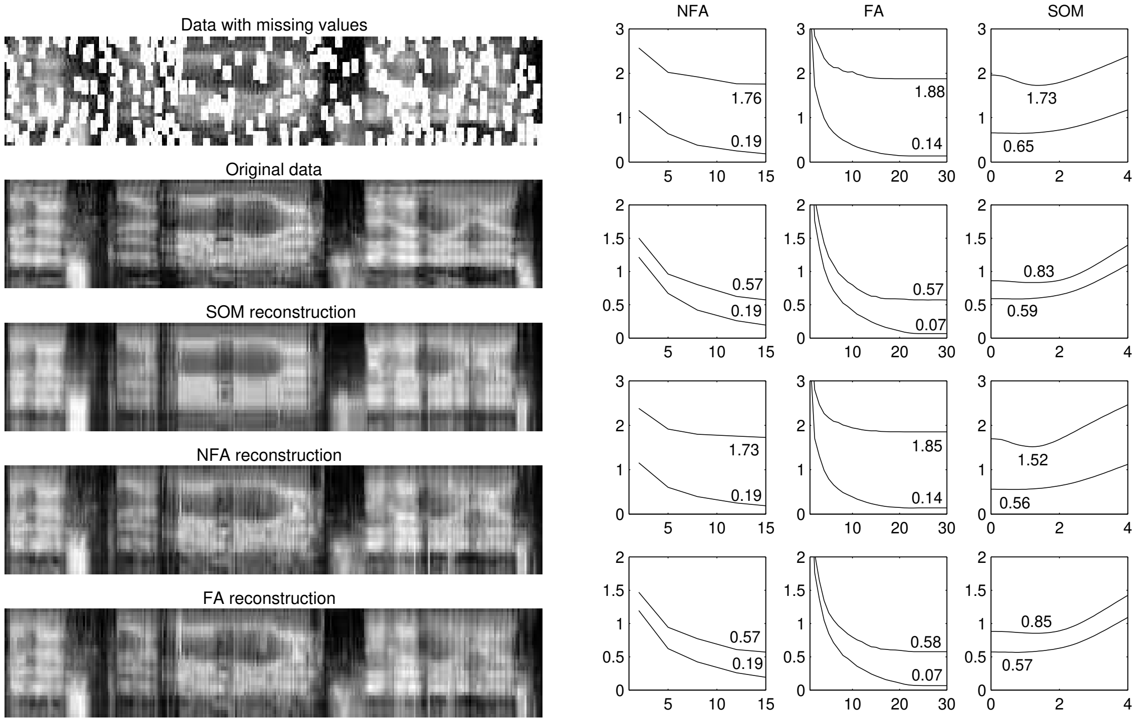


Figure 4.7: Left: Speech data reconstruction example with best parameters of each algorithm. Right: Mean square reconstruction errors of speech data as a function of number of factors or softening width. Reading from up to down, on the first and third row the values are missing in patches, and on the second and fourth, randomly. In the upper rows the test data is new words and in the lower rows the data is permuted randomly. Upper curve in each plot is the test result for missing values and lower is the reconstruction of observed values.

nonlinear factor analysis are closer to linear factor analysis than highly nonlinear mappings such as SOM. Nonlinear factor analysis is nevertheless able to capture nonlinear structure in the data and performed as well or better than linear factor analysis in all the reconstruction tasks.

Detection of process state changes

One potential application for the nonlinear dynamic state-space model discussed in Section 4.3 is process monitoring. In [1], 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. Figure 4.8 shows the observations on the left and the estimated underlying states on the right.

The first 1000 samples of the observations were used to learn the nonlinear dynamical system. The model was then fixed and applied to new observations. In the middle of the new data set, at time 1500, the underlying dynamics of one of the Lorenz processes abruptly changes. It is very difficult to detect this from the observed nonlinear mixtures whereas the estimated sources clearly exhibit the change. The cost function used in ensemble learning can readily be used for monitoring the change as can be seen in Figure 4.9. It

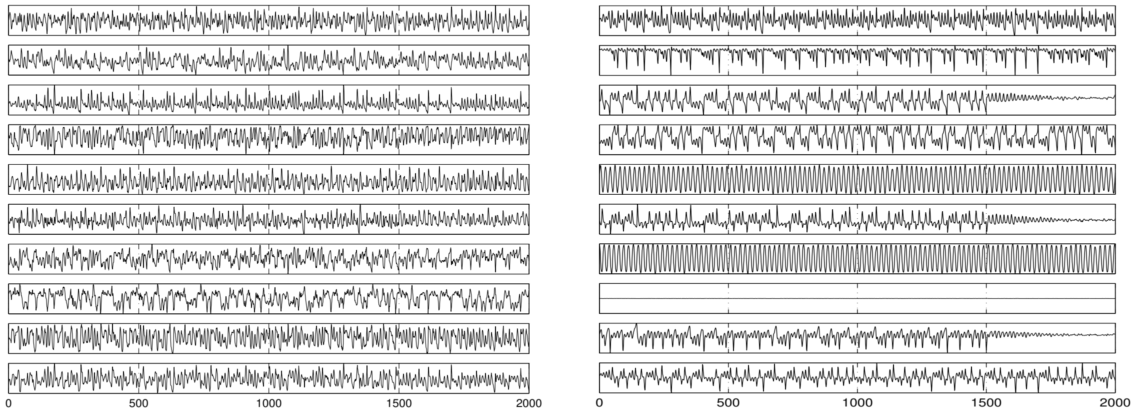


Figure 4.8: Observations on the left and reconstructed sources on the right. The process changes abruptly at time 1500 but this is visible only in the sources.

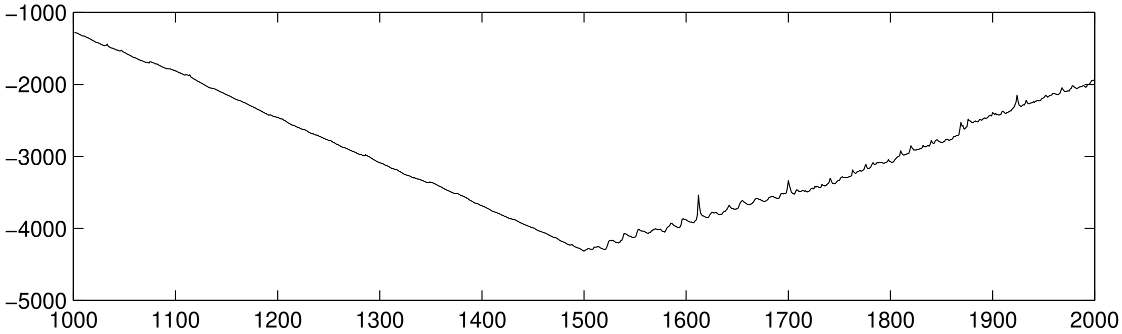


Figure 4.9: The change in the process can be detected by monitoring the cost function.

was shown experimentally that the method outperforms several standard change detection methods in this task [1].

References

[1] 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.

[2] T. Raiko and H. Valpola. Missing values in nonlinear factor analysis. In *Proc. ICONIP 2001*, pages 822–827, Sanghai,China, November 2001.

Chapter 5

Computational neuroscience

Aapo Hyvärinen, Ella Bingham, Mika Inki, Raju Karthikesh, Patrik Hoyer, Jarmo Hurri

5.1 Advances in statistical generative models

Aapo Hyvärinen, Ella Bingham, Mika Inki, Raju Karthikesh

In this project, various extensions to the theory of independent component analysis and blind source separation were developed.

First, we developed different methods of performing blind source separation using other information than that used by ordinary ICA. These new methods exploit the temporal structure of the signals.

The first method [1] models signals using nongaussian autoregressive models. As in basic ICA, nongaussianity is paramount here, but in this case it is not directly the nongaussianity of the signal, but of the noise (innovation) signal that drives the autoregressive process. This can also be understood as finding those projections of the data that are the most "interesting": Those projections that are the most structured are considered as the most interesting. This is reminiscent of the basic ICA estimation principle: Find those projections that are the most structured in the sense of being the most nongaussian.

A second method [2] in the same vein considers signals that have nonstationary variance. This is a very different form of time structure from the one above. We showed how such a structure can be interpreted simply using cumulants, which also provide an efficient algorithm in the same way as FastICA.

What these two methods have in common is the extension of the separation ability to gaussian signals. Whereas ordinary ICA can only separate nongaussian signals, these two methods can separate gaussian signals if other properties of the signals match those used in the underlying models. An example of a gaussian signal that can be separated by the latter method is given in Fig. 5.1.

All commonly used methods for ICA assume that the number of independent components is not larger than the number of observed signals. If there are more independent components, estimation becomes much more complicated. This case, called the case of "overcomplete bases" requires new estimation methods. Previously proposed methods are computationally very demanding, so we attempted to find algorithms that can do the computations needed for the estimation with reasonable computational load. The methods we found [3] are based on rather heuristic extensions of some well-known ICA estimation principles.

Also, we investigated a very interesting and surprising connection between ICA and the well-known nonlinear regression method of backpropagation using multi-layer perceptrons. We showed [4] that if the data is modelled by ICA, the theoretically optimal nonlinear regression is very closely approximated by a multi-layer preceptron whose weight vectors are given by the elements of the ICA mixing matrix

Finally, we developed a method [5] for using prior information in ICA estimation. We argued that the most interesting form of prior information that one can use in this context is the information on the sparseness of the mixing matrix, i.e. most elements of the mixing matrix are practically zero. Such a prior can easily be incorporated in ICA estimation because it is a conjugate prior, thus no new algorithms need to be developed because the prior information can be expressed in the form of a virtual sample, i.e. addition of artificial data to the sample.

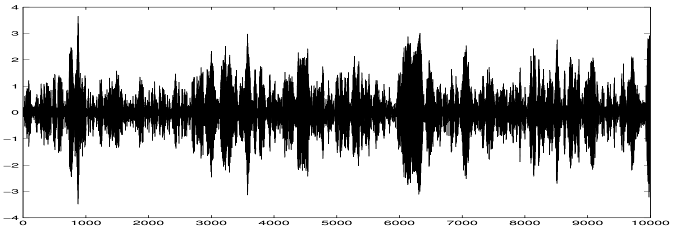


Figure 5.1: A signal that is gaussian, i.e. cannot be separated by ICA from linear mixtures. Our new method [2] can separate this kind of signals, however.

References

[1] A. Hyvärinen. Complexity pursuit: Separating interesting components from time-series. *Neural Computation*, 13(4):883–898, 2001.

[2] A. Hyvärinen. Blind source separation by nonstationarity of variance: A cumulant-based approach. *IEEE Trans. on Neural Networks*, 12(6):1471–1474, 2001.

[3] M. Inki and A. Hyvärinen. Two methods for estimating overcomplete independent component bases. In *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA2001)*, San Diego, California, 2001.

[4] A. Hyvärinen and E. Bingham. Connection between multilayer perceptrons and regression using independent component analysis. *Neurocomputing*. in press.

[5] A. Hyvärinen and R. Karthikesh. Imposing sparsity on the mixing matrix in independent component analysis. *Neurocomputing*. in press.

5.2 Statistical models of visual processing

Patrik Hoyer, Jarmo Hurri, Aapo Hyvärinen

Our research concentrates on modelling visual perception using statistical models. We 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. 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 independent component analysis (ICA), a statistical technique that has recently gained widespread attention in neurocomputing and signal processing. The goal is to transform a data vector into components that are statistically independent. In [1,9] it was shown that ICA applied to natural image input yielded features closely resembling the spatial receptive fields of simple cells in the mammalian primary visual cortex. We have extended this result by showing that ICA can also explain chromatic and binocular properties of simple cell receptive fields. This was done by applying the method to natural colour and stereo images and comparing the learned features (see Figure 5.2) with known receptive field properties [3].

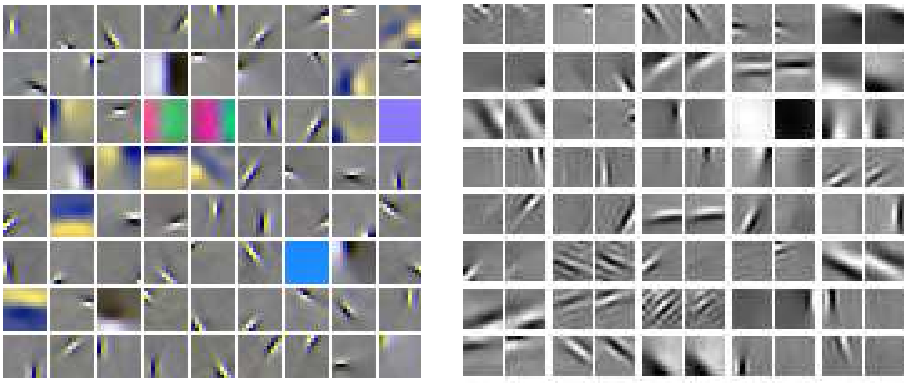


Figure 5.2: ICA features from colour and stereo image data. Left, a small subset of features estimated from colour images. Right, features from stereo images (each pair of patches depicts one binocular feature).

We have also developed extensions of ICA to model further properties of the primary visual cortex. Specifically, we wanted to additionally explain the response properties complex cells (the second major functional class of cells in the primary visual cortex) as well as the columnar structure of the cortex. For this purpose, we have developed a model called *topographic independent component analysis* [8], that applied to natural image data yields both simple and complex cell receptive fields and a topography like that observed in the visual cortex [6,7]. Figure 5.3 shows an estimated basis for image data.

Can further properties of the visual system be explained by this information-theoretic approach? We have considered how the responses (to natural images) of complex cells could be sparsely represented by a higher-order neural layer [2]. Essentially, we filtered our images with a widely accepted complex-cell model (see Figure 5.4) and performed ICA on this data. This leads to an interesting kind of contour coding, where units are selective to the contour length. A subset of the learned components is shown in Figure 5.5. In addition, the model can explain certain extra-classical receptive field effects of neurons in the visual cortex.

We have also started to explore the statistical structure of natural image sequences,

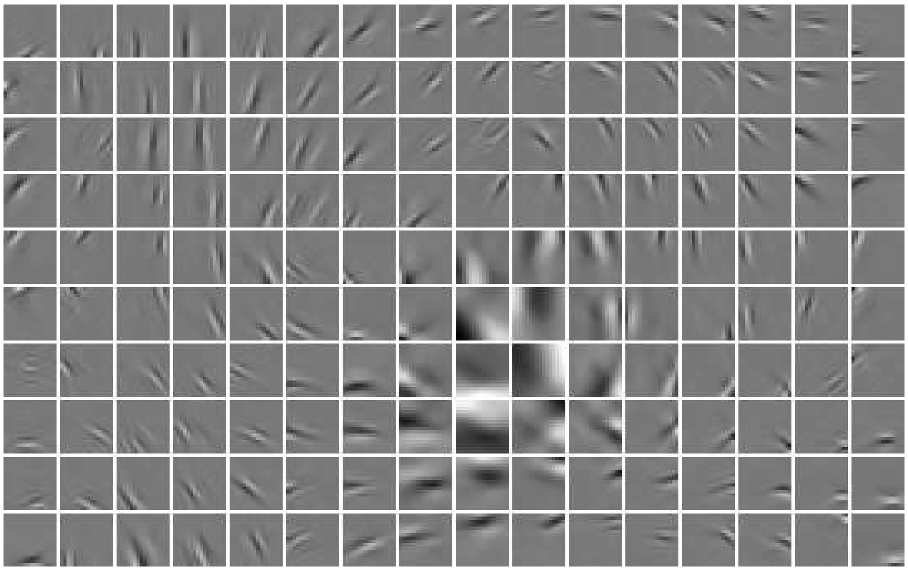


Figure 5.3: Topographic independent component basis estimated from natural image data. Each patch corresponds to the receptive field of a simple cell, whereas the responses of the model complex cells are given by locally pooled responses of the simple cells. Note that neighboring units tend to have similar location, orientation and spatial scale, making the topography very similar to the one found in the visual cortex.

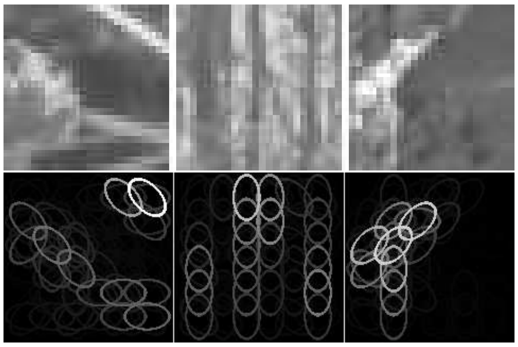


Figure 5.4: Illustration of the complex-cell responses to natural image patches. Top row: Three patches from our set of natural images. Bottom row: Corresponding responses of the model complex cells to the patches. The ellipses show the orientation and approximate extent of the individual complex cells. The brightness of the different ellipses indicate the response strengths.

that is, natural video. We have developed a simple model of the temporal structure of linear filter outputs based on a nonlinear form of autocorrelation. This approach leads to the emergence of simple cell properties as well [4]. Thus, we have an alternative approach to sparse coding or independent component analysis. The advantages of this alternative approach are a problem that we will investigate in the future.

In addition to these experiments regarding natural image statistics, we have also investigated the theoretical foundations of independent component analysis, as related to the neurobiological principle of maximum information transfer (infomax). In [5] it was shown that the standard formulation of nonlinear infomax is very sensitive to the assumed nonlinearity, but that an alternative formulation based on a non-Gaussian noise model gives similar predictions but is much more robust.

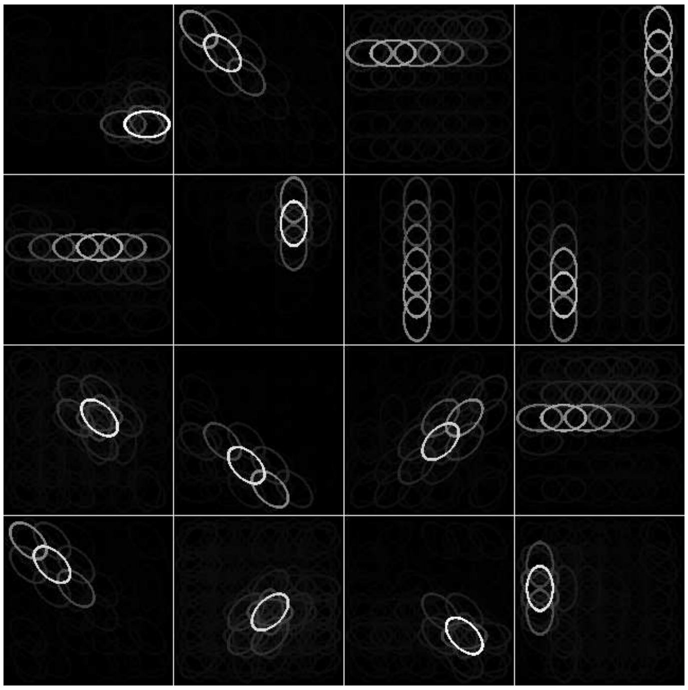


Figure 5.5: A subset of the learned higher-order features. The features code for *collinear* active complex cells, representing contours of varying length in the images.

References

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

[2] P. O. Hoyer and A. Hyvärinen. A multi-layer sparse coding network learns contour coding from natural images. *Vision Research*. In press.

[3] P. O. Hoyer and A. Hyvärinen. Independent component analysis applied to feature extraction from colour and stereo images. *Network: Computation in Neural Systems*, 11(3):191–210, 2000.

[4] J. Hurri and A. Hyvärinen. Simple-cell-like receptive fields maximize temporal coherence in natural video. Submitted.

[5] A. Hyvärinen. An alternative approach to infomax and independent component analysis. *Neurocomputing*. In press.

[6] A. Hyvärinen and P. O. Hoyer. Emergence of topography and complex cell properties from natural images using extensions of ICA. In *Advances in Neural Information Processing Systems*, volume 12, pages 827–833. MIT Press, 2000.

[7] A. Hyvärinen and P. O. Hoyer. A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images. *Vision Research*, 41(18):2413–2423, 2001.

[8] A. Hyvärinen, P. O. Hoyer, and M. Inki. Topographic independent component analysis. *Neural Computation*, 13(7):1527–1558, 2001.

[9] B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

Chapter 6

Biomedical data analysis

Ricardo Vigário, Jaakko Särelä, Harri Valpola, Erkki Oja

6.1 Introduction

In a combination of expert efforts from the Laboratory of Computer and Information Science and the Brain Research Unit (both from the Helsinki University of Technology), we have, over the past 5 years, introduced ICA as a tool for the analysis of electroencephalographic (EEG) and magnetoencephalographic (MEG) recordings. In the early years we have shown that ICA is very suited for the identification and removal of artifacts, as well as the decomposition of several types of evoked brain responses.

In the period spanned by this report, we have started to introduce further information in the search for meaningful signal components. This has, at times, lead to relaxing, but not rejecting, the independence assumptions.

Additionally, we have started to enlarge our areas of application to biomedical image processing.

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

Validation of the linear ICA model for EEG and MEG

The application of ICA to the study of EEG and MEG signals assumes that several conditions are verified: the existence of statistically independent source signals, their instantaneous linear mixing at the sensors, and the stationarity of both the source signals and the mixing process.

In [1], making use of very few and plausible assumptions, we validate the use of ICA for the analysis of EEG and MEG recordings.

References

[1] R. Vigário, J. Särelä, V. Jousmäki, M. Hämäläinen, and E. Oja. Independent component approach to the analysis of EEG and MEG recordings. IEEE Trans. Biomed. Eng., 47:589–593, 2000.

6.2 Dipole modeled ICA decomposition of evoked responses

When applied to the decomposition of evoked responses, independent component analysis is often used in a blind source separation framework, *i.e.*, the goal is to identify underlying source signals from their instantaneous linear mixtures, with no prior knowledge on the sources or the mixing process. Yet, often some modeling is explicitly or implicitly added when attempting to give a physiological plausible interpretation to the result of the decomposition.

In [1], we propose a dipole modeled ICA decomposition (see [2], for better explanation). There, we assume that all the relevant independent components can be modeled using equivalent current dipoles. The incorporation of this modeling information simplifies the search for the underlying brain activated sources, while giving better physiological grounds to ulterior interpretations of the decomposed signals.

Figure 6.1 shows the complete set of 122 averaged magnetic evoked responses to combined auditory and somatosensory stimulation. Four inserts in the figure highlight equal number of interesting signal types present in the measurements: IV corresponds to step-like signal; III, a sharp and early response, originates in the primary somatosensory cortex; II and I are broader signals, with longer latencies than the previous one, and are associated with the primary auditory responses from the right and left hemispheres.

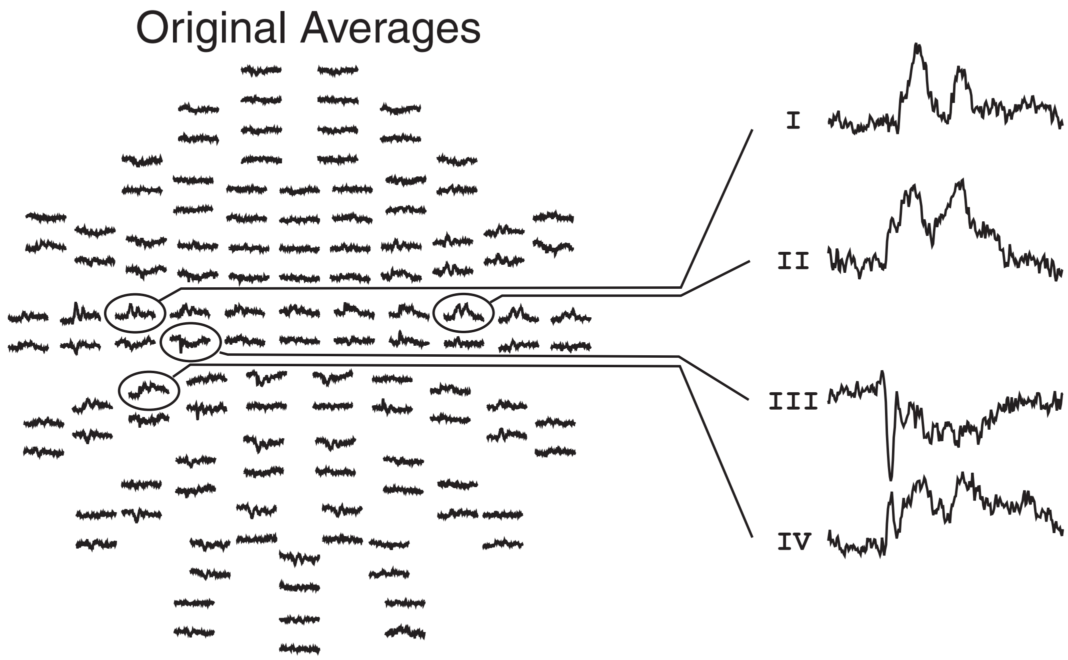


Figure 6.1: Average evoked MEG recordings to concomitant vibrotactile and auditory stimulation. Onsets show the maximum responses for the stimulation.

Both the original (Or) and the dipole modeled (Dm) ICA approaches give good results in the decomposition of the evoked responses, as can be seen, in Fig. 6.2, for DmICA. Note that each component corresponds to only the auditory or the somatosensory brain responses. Yet, we can see as well that the newly introduced algorithm requires a much simpler neural modeling to justify each one of its components (c.f. [2] for details on the goodness of fit values). In Fig. 6.3a) are plotted the same 4 signals seen in Fig. 6.1. Underneath are the remaining of these signals once the first, second, third and fourth original independent components are removed. Fig. 6.3b) present the same type of results for the

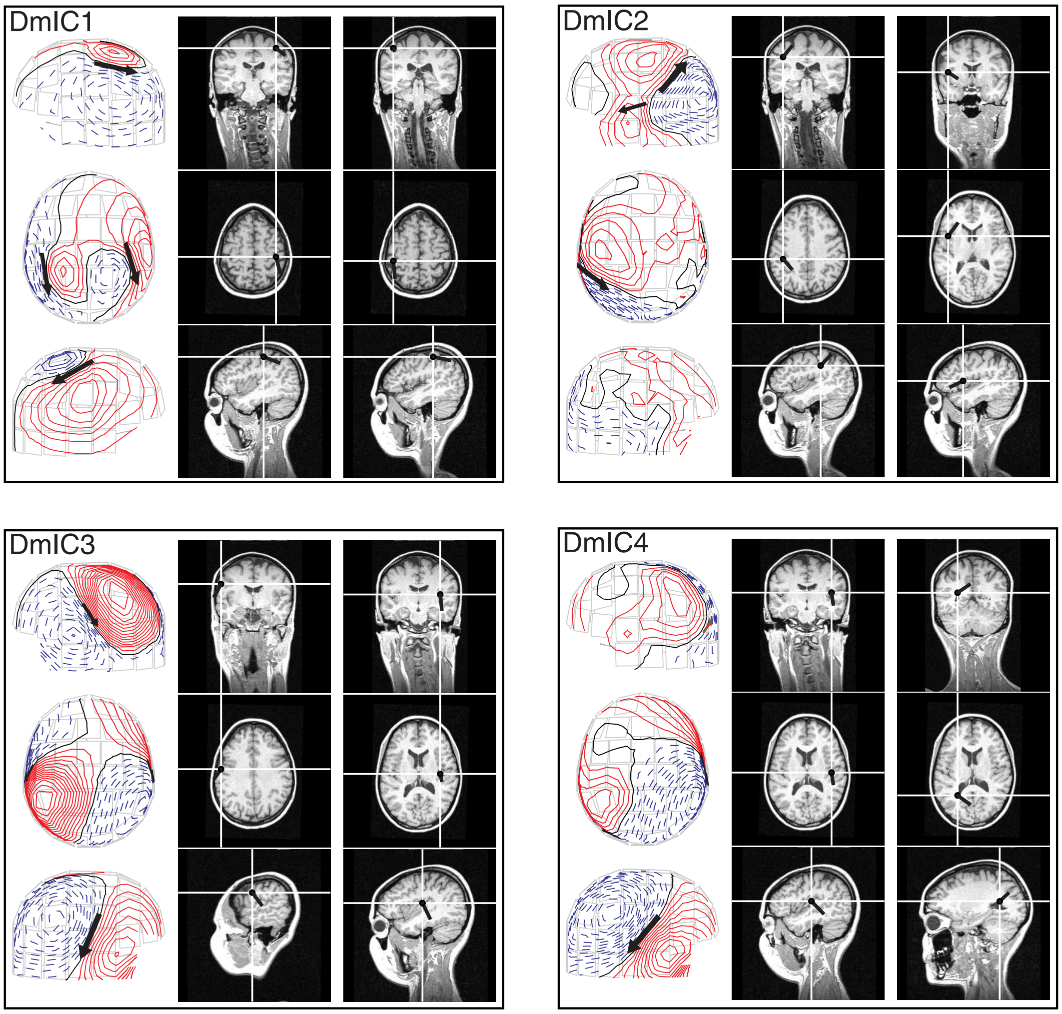


Figure 6.2: Field patterns and localizations of the independent components found by the dipole modeled ICA method.

dipole modeled algorithm. In OrICA, the extraction of the first component removes all auditory responses, whereas DmICA discriminates between the left and the right responses, after removal of the third and fourth components, respectively.

References

[1] R. N. Vigário. Dipole modeling in FastICA decomposition of evoked responses. Proceedings of the 2nd Int. Workshop on Independent Component Analysis and Blind Separation of Signals, ICA'00 (Helsinki, Finland), 2000.

[2] R. Vigário and E. Oja. Independence: a new criterion for the analysis of the electromagnetic fields in the global brain? Neural Networks, 13:891–907, 2000.

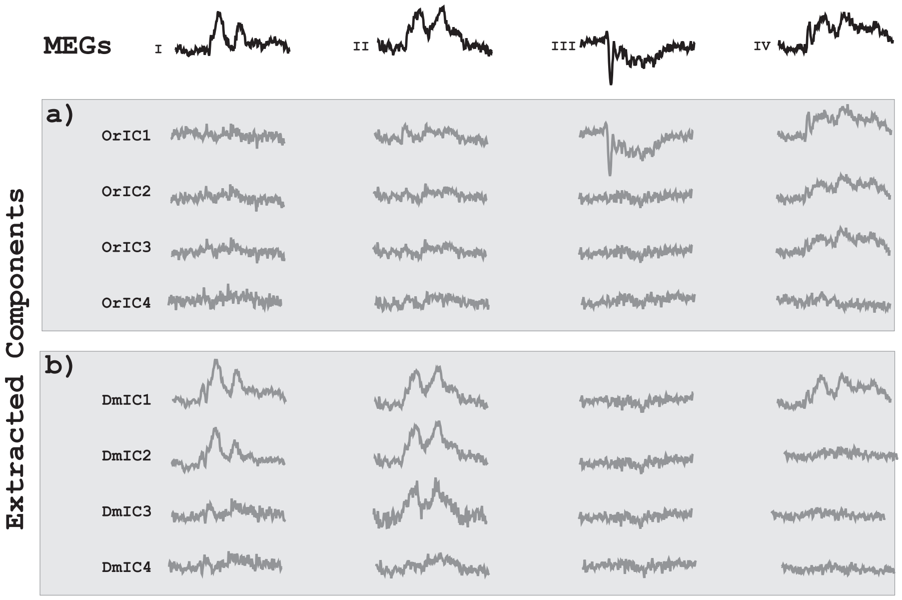


Figure 6.3: Original four MEG averages (see Fig. 6.1), and the result of extracting each of the first four independent components, using the original ICA algorithm (a) or the dipole modeled one (b).

6.3 Analysis of rhythmic electromagnetical brain activity

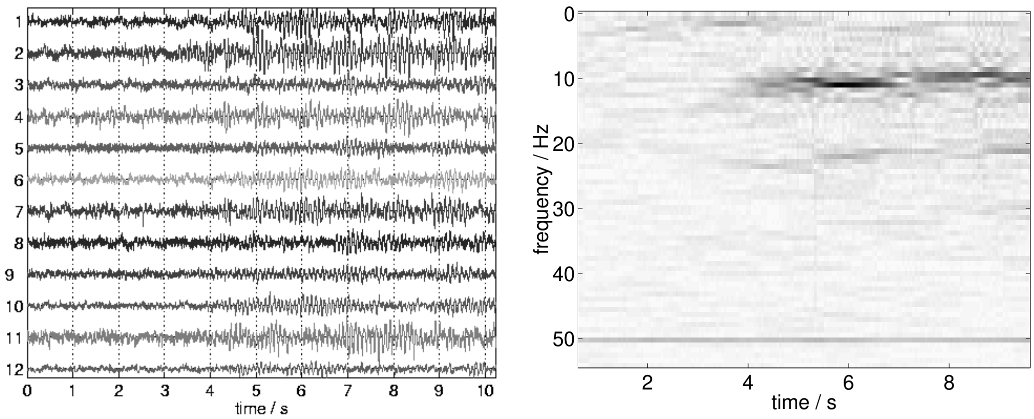
Periodicity information in extraction of cardiac artefacts

In [1], we propose an algorithm, based on the deflationary version of the FastICA, which uses prior information to find the periodic and quasi-periodic signals of interest out of a number of MEG recordings. This algorithm joins some advantages of high-order statistics methods to the ones of the temporally based ones. It has been used to extract, very accurately, the cardiac artifact present in some MEG data, without the need for any additional electrical measurements

Dynamical factor analysis of rhythmic MEG signals

In dynamical factor analysis (DFA) [2], the dynamics of the underlying sources are modelled using a state-space model. The state transitions themselves are modelled by nonlinear feedforward networks with one hidden layer. The underlying sources are observed through a linear process as in ICA. DFA is learned using *ensemble learning*, which is a full distribution approximation of correct Bayesian treatment (for a more detailed description see Chapter 4).

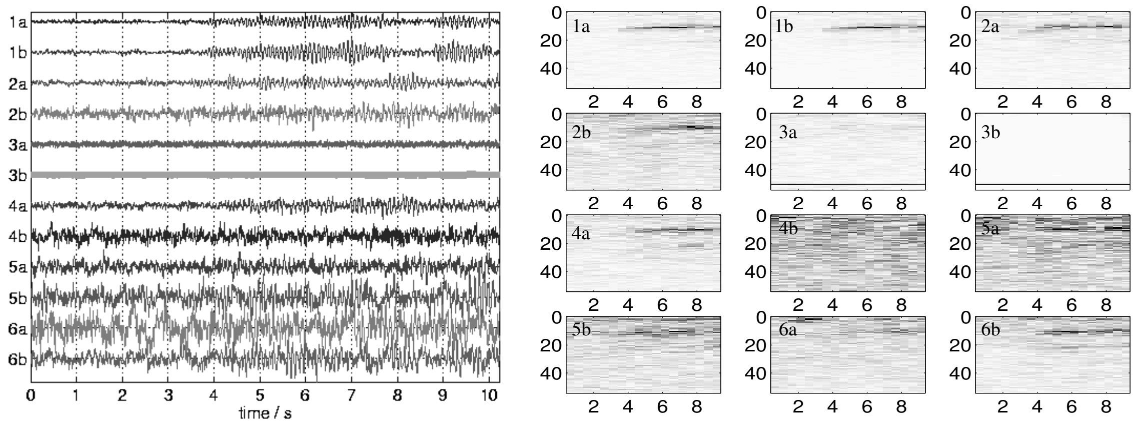
DFA can model any dynamics, given enough units in the hidden layer, but the model is at its best when modelling smooth, close to linear dynamics. In [2] DFA is used for analysis of rhythmic activity in MEG. Typically, these rhythms contain a main frequency, but also some additional harmonics, making their dynamics non-linear, but close to linear. Fig. 6.4 (a) shows a short period of time of the recordings on twelve channels and (b) shows a portion of the short time fourier transform of the first MEG channel. From both figures it is evident that there exist α-activity, characterised by a frequency around 10Hz, on the latter half, but the channels are quite noisy.



(a) (b)

Figure 6.4: (a) Short fragment of MEG recording over twelve channels. (b) Short time Fourier transform of the first channel of MEG recordings. α-rhythm is visible after 4s.

The resulting DFA decomposition, shown in Fig. 6.5, has been able to separate several α-generators (*e.g.* sources 1a, 1b, 2a, 2b and 4a) with very clear activations around 10Hz, but also some activation on the first harmonics around 20Hz, suggesting non-linear dynamics. These sources are less noisier than the sources found by traditional ICA methods using higher order statistical information or temporal correlation information. DFA also perfectly separates both phases of the 50Hz power line (sources 3a and 3b), which are both present in the measurements.



(a) (b)

Figure 6.5: (a) Resulting decomposition. (b) Short time Fourier transform of the decomposition.

DFA performs an explicit modeling of the sources. The dynamical part of DFA can determine, at each realization t of the observation $\mathbf{x}(t)$, which portion corresponds to the underlying estimated sources, $\mathbf{s}(t)$, and which corresponds to noise, $\mathbf{n}(t)$. DFA can therefore deal better with noisy signals than algorithms that are uncapable of such explicit modeling, even though they may use implicitly some temporal information.

Ensemble learning favors simpler and smoother models. Furthermore, the predictions are made over a collection of models. The probability of overfitting to a particularly strong, but very peaky, posterior is therefore very small. Because DFA uses ensemble learning to estimate the models for the factors, these are less prone to overlearning (see Section 2.2 for discussion of overlearning in ICA).

References

[1] Allan Kardec Barros, Ricardo Vigário, Veikko Jousmäki, and Noboru Ohnishi. Extraction of periodic signals from multi-channel bioelectrical measurements. IEEE Trans. Biomed. Eng., 47:583–58, 2000.

[2] J. Särelä, H. Valpola, R. Vigário, and E. Oja. Dynamical factor analysis of rhythmic magnetoencephalographic activity. Proceedings of the 3rd international workshop on independent component analysis and blind signal separation, ICA'01, (San Diego, CA, United States), 2001. Paper number 44.

Chapter 7

Image analysis applications

Erkki Oja, Jorma Laaksonen, Jukka Iivarinen, Markus Peura, Markus Koskela, Sami Laakso, Jussi Pakkanen, Ville Viitaniemi

7.1 Content-based image retrieval by Self-Organizing Maps

Erkki Oja, Jorma Laaksonen, Markus Koskela, Sami Laakso, Ville Viitaniemi

Content-based image retrieval (CBIR) has been a subject of very intensive research effort for more than a decade. 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

We have developed a neural-network-based CBIR system named PicSOM [1, 2]. The name stems from "picture" and the Self-Organizing Map. The SOM is used for unsupervised and topology-preserving mapping from the image descriptor space to a two-dimensional lattice, or grid, of artificial neural units. The PicSOM system is built upon two fundamental principles of CBIR: query by pictorial example and relevance feedback [3].

The methodological novelty of PicSOM 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. Every image query is unique and each user of a CBIR system has her own transient view of image similarity and relevance. Therefore, a system structure capable of holding many simultaneous similarity representations can adapt to different kinds of retrieval tasks. In the PicSOM approach, the system is able to discover those of the parallel Self-Organizing Maps 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) [4]. 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). This makes the time complexity of the search $O(n)$, where n is the number of SOM units. With the TS-SOM one can, however, follow the hierarchical structure and reduce the complexity of the search to $O(\log n)$. This reduction can be achieved by first training a smaller SOM and then creating a larger one below it so that the search for the BMU on the larger map is always restricted to a fixed centered below the already-found BMU on the above map.

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. This is illustrated in Figure 7.1, where MPEG-7 *Edge Histogram* descriptor has been used as the feature. It can be seen that, e.g., there are many ships in the top-left corner of the map surface, standing people and dolls beside the ships, and buildings in the bottom-left corner. Visually – and also semantically – similar images have thus been mapped near each other on the map.

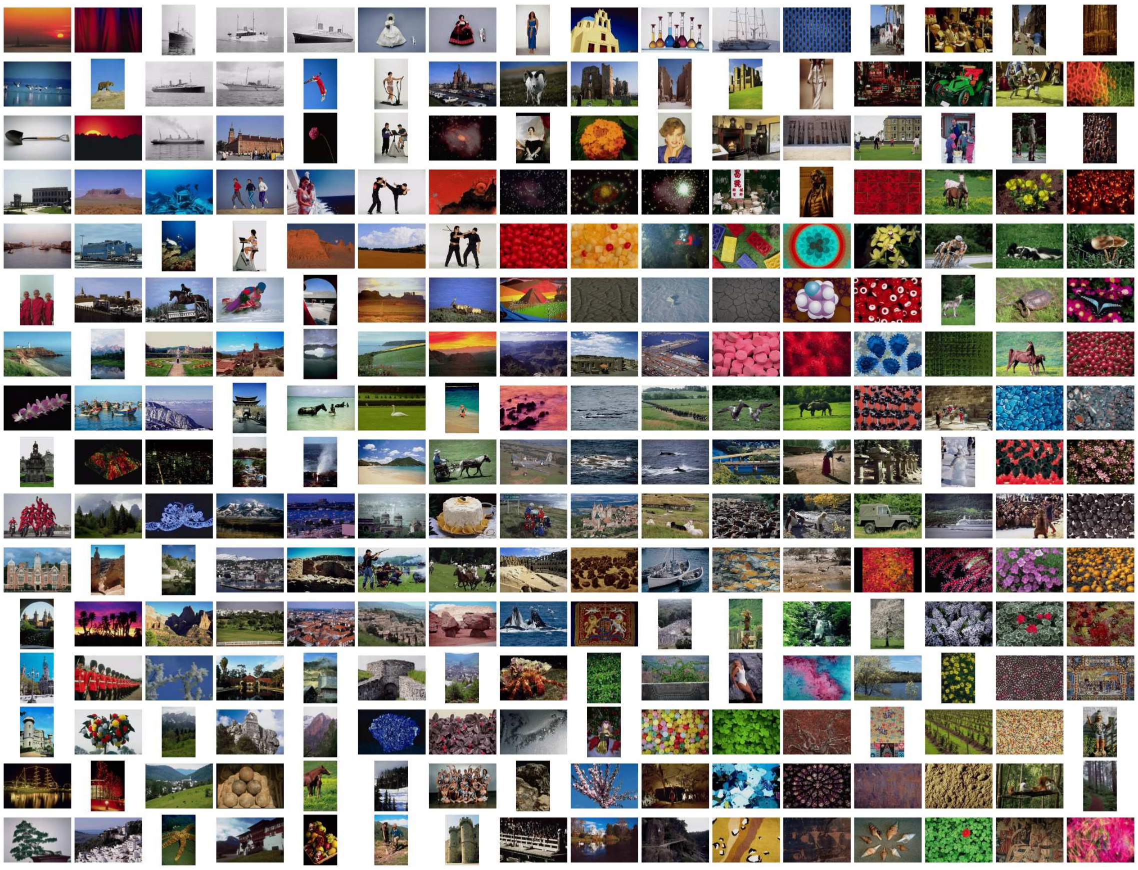


Figure 7.1: The surface of a 16×16-sized TS-SOM level trained with the MPEG-7 *Edge Histogram* descriptor.

Self-organizing relevance feedback

Each image seen by the user of the system is graded by her as either relevant or irrelevant. All these images and their associated relevance grades are then projected on all the SOM surfaces. This process forms on the maps areas where there are 1) many relevant images mapped in same or nearby SOM units, or 2) relevant and irrelevant images mixed, or 3) only irrelevant images, or 4) no graded images at all. Of the above cases, 1) and 3) indicate that the corresponding content descriptor agrees well with the user's conception on the relevance of the images. Whereas, case 2) is an indication that the content descriptor cannot distinguish between relevant and irrelevant images.

When we assume that similar images are located near each other on the SOM surfaces, we are motivated to spread the relevance information placed in the SOM units also to the neighboring 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, irrelevant images receive negative weights that are inversely proportional to the number of irrelevant images. The overall sum of these relevance values is thus zero. The values are then summed in the BMUs of the images and the resulting sparse value fields are low-pass filtered. Figure 7.2 illustrates how the positive and negative responses, displayed with red and blue map units, respectively, are first mapped on a SOM surface and how the responses are expanded in the convolution. Content descriptors that fail to coincide with the user's conceptions produce

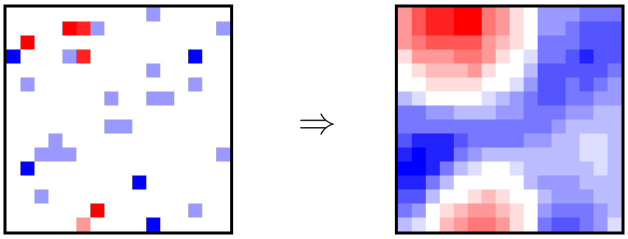


Figure 7.2: An example of how a SOM surface, on which the images selected and rejected by the user are shown with red and blue marks, is convolved with a low-pass filter.

lower qualification values than those descriptors that match the user's expectations. As a consequence, the different content descriptors do not need to be explicitly weighted as the system automatically takes care of weighting their opinions.

After removing duplicate images, the second stage of processing is carried out. Now, the qualification values of all images in this combined set are summed up on all used SOMs to obtain the final qualification values for these images. Then, a set of images with the highest qualification values are returned as the result of the query round.

MPEG-7 content descriptors

Development of content-based image retrieval techniques has suffered from the lack of standardized ways for describing visual image content. Luckily, MPEG-7 international standard is now emerging 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 [5] and shown that they can be successfully used with it.

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, S. Laakso, and E. Oja. Self-Organizing Maps as a relevance feedback technique in content-based image retrieval. *Pattern Analysis and Applications*, 4(2+3):140-152, June 2001.

[3] G. Salton and M. J. McGill Introduction to Modern Information Retrieval. McGraw-Hill, 1983.

[4] 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.

[5] J. Laaksonen, M. Koskela, and E. Oja. PicSOM – Self-Organizing Image Retrieval with MPEG-7 Content Descriptors. Accepted to *IEEE Transactions on Neural Networks*, 2002.

7.2 Fault analysis of running paper web

Jukka Iivarinen

On-line inspection is an essential part of modern web or sheet manufacturing. The purpose of an inspection system is to detect and classify those defects which impair the quality of a product as compared to the requirements set by a user. Typical characteristics of web manufacturing processes are the large values of web width and production speed. The web width of a modern paper machine may exceed 9 metres and its speed may reach 30 m/s. Such a machine makes about $3{\cdot}10^8\text{mm}^2$ of paper each second, and all that production has to be inspected with a 100 % coverage and with a resolution even less than 1 mm^2.

Historically defect detection of web surfaces has been accomplished by hardware solutions for thresholding and matched filters. These techniques have made possible to detect only the most basic defect types. Detection of more complicated but critical defect types has remained unreliable. In this project more sophisticated methods have been developed. Use of texture and more complicated classifiers have become possible due to new sensor technology, increased calculation capability of computers and specialized hardware. Surface inspection has been studied in our laboratory since 1995 [1, 2]. The main interest has been the detection and classification of defects in a running paper web.

Overview of the method

The proposed system model for web inspection has two phases: a segmentation phase (defect detection) and a classification phase (defect classification) (Figure 7.3). In the segmentation phase feature extraction is done and potential defect areas are marked. In the classification phase features describing the shape and internal structure of defects are extracted and defects are classified to different defect classes.

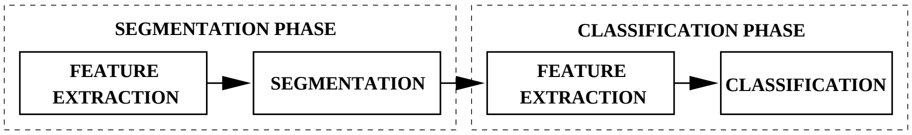


Figure 7.3: The proposed method for web inspection consists of two phases.

Segmentation phase A modified self-organizing map, called the statistical SOM, is used to separate defects from a fault-free web. A set of co-occurrence matrix features is used for the texture description. The feature distribution of fault-free samples is estimated with the statistical SOM. Now fault detection is based on the following idea: an unknown sample is classified as a defect if it differs enough from this estimated distribution. The segmentation scheme is depicted in Figure 7.4.

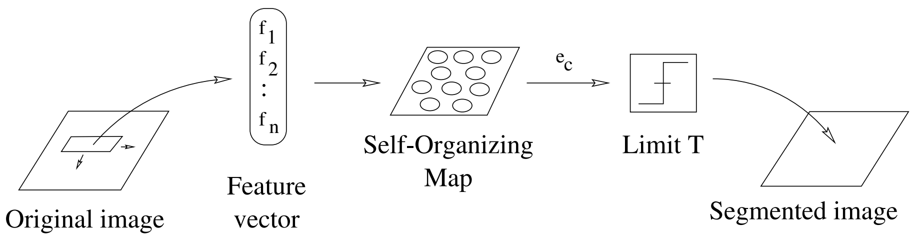


Figure 7.4: The segmentation scheme.

Classification phase In defect classification the SOM is used to cluster unknown defects. It finds the classes (clusters) that are inherent to defect samples. These classes are then given an explanation by hand. Classification is based on the visual appearance of defects. Five simple shape descriptors are used to characterize the shape, and a gray level histogram and some co-occurrence matrix features are used for the internal structure.

An example defect and its features are depicted in Figure 7.5(a). In Figure 7.5(b)-(d) are the classification results. Each column shows the best-matching defect when using different features. In the bottom row are the numbers and the weight vectors of the best-matching SOM units. The first and the second row show one of the typical defects that belongs to the best-matching unit. The labels of the best-matching units are *elongated, smooth shape* (SSD), *light spot* (GLH), and *light spot* (TEX). The final classification is then *elongated, smooth, light spot*.

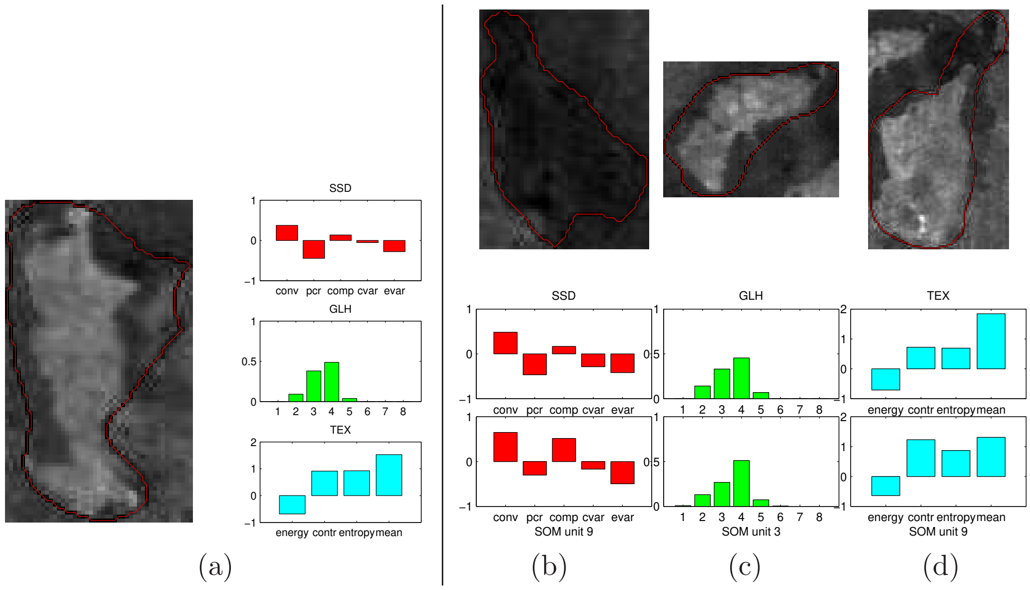


Figure 7.5: (a) A defect and its features. (b)-(d) The best matching defects to the defect in (a) with respect to different features.

Conclusions

The two-stage approach offers advantages when considering real-time processes. Speed requirements of real-time defect detection and classification can be satisfied by splitting the procedure into two stages. The defect detection is a suitable part to hardware implementation, whereas more time can be spend on classifying found defects. In defect detection the classifier is taught with examples of fault-free surface while in defect classification shape and internal structure characteristics of defects are learned from examples.

References

[1] J Iivarinen. Unsupervised Segmentation of Defect Images. In D. P. Casasent, Ernest L. Hall, editors, *Intelligent Robots and Computer Vision XX: Algorithms, Techniques, and Active Vision*, Proc. SPIE 4572, pp. 488–495, 2001.

[2] J. Iivarinen, K. Heikkinen, J. Rauhamaa, P. Vuorimaa, and A. Visa. A defect detection scheme for web surface inspection. *International Journal of Pattern Recognition and Artificial Intelligence*, 14(6):735–755, Sept. 2000.

7.3 Content-based retrieval of defect images

Jussi Pakkanen and 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 a database 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. The project started on March 2001.

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. There exist several defect classes (e.g. dark and light spots, holes, and wrinkles) that are fuzzy and overlapping, so it is not possible to label defects unambiguously.

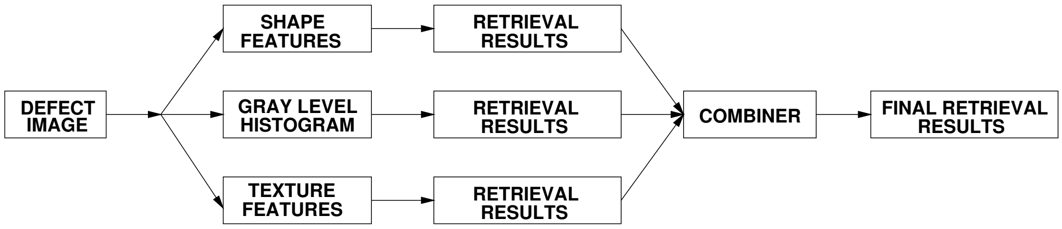


Figure 7.6: The structure of a CBIR system for defect images.

In the present work we have adopted the PicSOM system as our content-based image retrieval (CBIR) system and embedded a defect image database into PicSOM. PicSOM has several features that make it a good choice for our purposes. The basic structure of our defect PicSOM system is shown in Figure 7.6. Given a new defect image, three different feature sets are extracted, similar defects according to each feature set are retrieved from the database, and these retrieved images are combined to produce the final, combined set of similar defect images.

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.

Six simple shape descriptors are used to characterize the shape of a defect. The descriptors are: convexity, principal axis ratio, compactness, circular variance, elliptic variance, and angle. The internal structure can be regarded as a distribution of gray levels in a defect. This distribution can be characterized as a simple gray-level histogram. In addition, four texture features, energy, contrast, entropy, and mean, calculated from the co-occurrence matrix are also used.

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 defect database has almost 13000 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 Industry Oy.

The example query images in Figure 7.7 show that the system works quite well. Under the TS-SOMs are the images selected by the user (or the query images), and on 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.

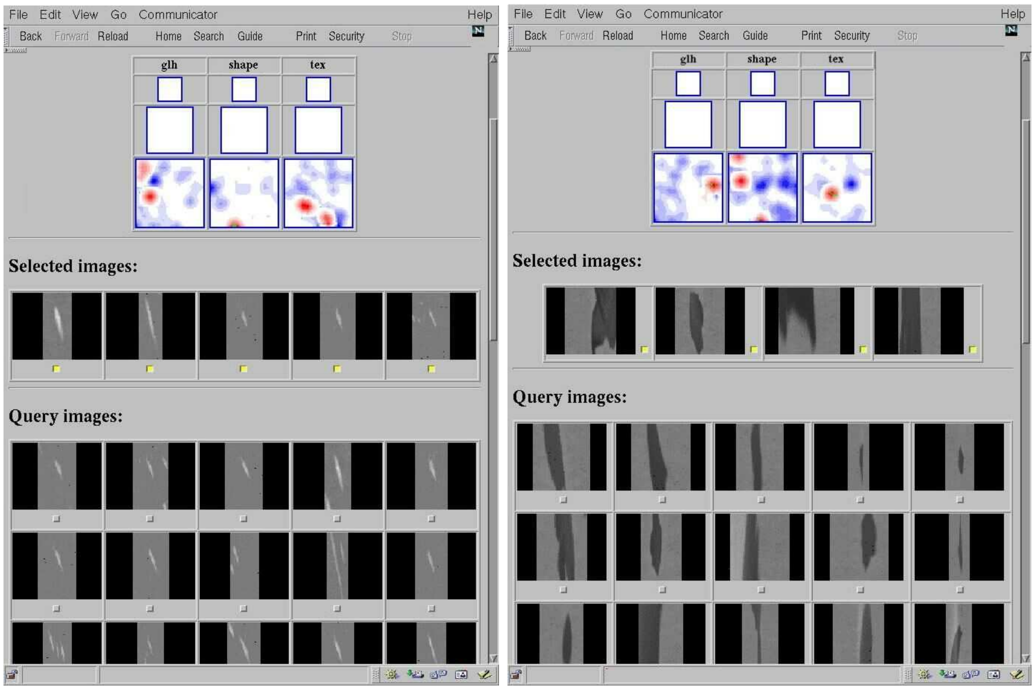


Figure 7.7: Some example PicSOM queries.

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 13000 surface defect images show that the system works fast with good retrieval results.

7.4 Attribute trees in image analysis

Markus Peura

Introduction

The appearance of many natural objects is irregular and indefinite. Spatial hierachy and dynamical behaviour are often additional challenges in the tasks of computer vision: automated recognition, classification, tracking, and prediction. Irregular objects are encountered frequently in biology, medical sciences, meteorology, and geosciences (Fig. 7.8).

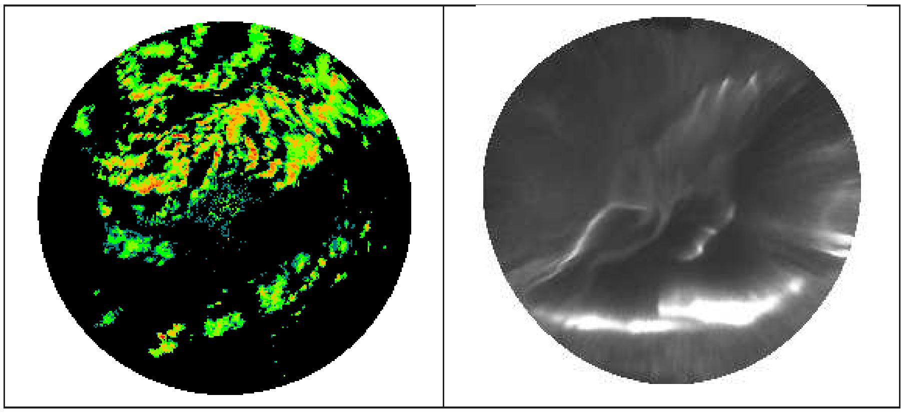


Figure 7.8: Examples of irregular natural images: precipitative clouds detected by a weather radar (left) and auroral lights captured by an all-sky camera (right).

Basic idea

The intensities of precipitation (rain) in a radar image can be interpreted analogously to terrain elevation on a topograhic map. The spatial hierarchy of the intensity segments can be extracted as a *tree*. The tree is a presentation of a topology and thus rotation, translation and scaling invariant. If nodes (or edges) are attached additional data - such as intensities, sizes, shape descriptions, coordinates, or types of objects - an *attribute tree* is obtained. An attribute tree is a compact presentation of an image as it preserves a significant part of the visual information and still has a storage size equal to a fraction of that required by the original image. Consequently, attribute trees can be processed faster than the original images.

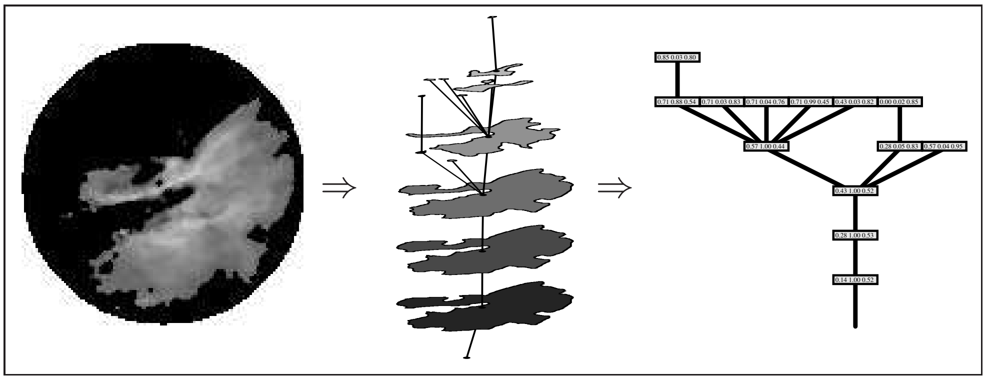


Figure 7.9: An image, its topology and attribute tree.

Fast tree matching

One of the central innovations in this study has been *the heuristic matching scheme for unordered attribute trees.* Denoting the size of a tree by N and the maximal out-degree by d, the proposed matching scheme has the complexity of $O(Nd \log d)$ while the exact algorithms are of $O(N^3)$. The obtained speed is essential in real time applications, like in *weather nowcasting* based on radar data.

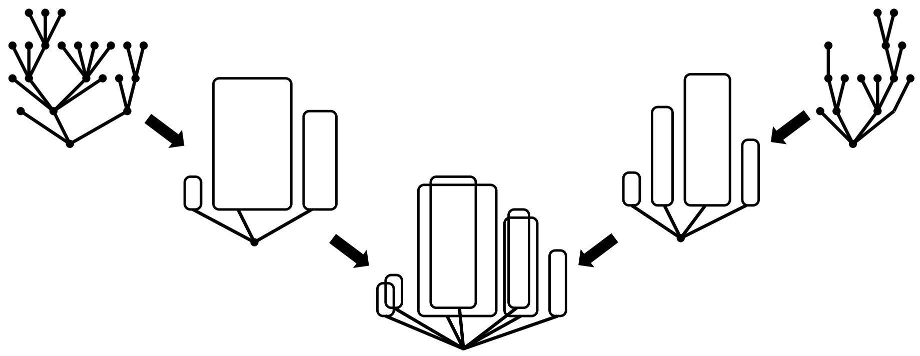


Figure 7.10: Illustration of index-based heuristic matching.

The scheme is based on dividing trees recursively into subtrees. The subtrees are matched according to indices, which have been calculated in advance using cumulative updating rules. In other words, exhaustive matching of subtrees is replaced by matching points in space, yielding an approximate matching result. The descriptors used in this study are height, child count (out-degree), node count (tree size), vertical centroid, and branching variance. If the node count is regarded as the mass of a tree, the centroid reflects the vertical distribution of the mass. The branching variance measures structural irregularity. A step of the matching scheme is illustrated in Fig. 7.10, where height and node count have been depicted as the height and width of the rectangles, respectively. When matching attribute trees, indices are calculated for attributes in similar fashion.

In comparison with the exact algorithms, the proposed linear-time approximation performs reliably with trees of up to a hundred nodes (Fig. 7.11). The proposed heuristic approach seems to outperform the exact algorithms even further when the quality criterion is the absolute number of matched nodes per computation time.

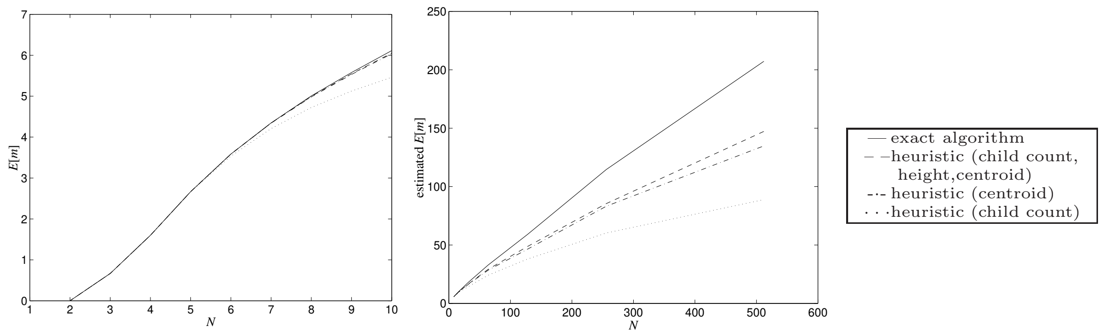


Figure 7.11: Matches found for non-isomorphic trees of size 1, 2, ..., 10 (complete sets) and 8, 16, ..., 512 (randomly sampled sets).

The Self-Organizing Map of attribute trees

After matching two trees, the trees can be merged in a *continuous fashion.* Such mixtures of matched trees are essentially interpolations and have applicability in learning systems that typically involve template generation through averaging. The *self-organizing map of attribute trees* [1, 2, 3] is an extension of the standard self-organizing map (applying vectors); the key issues is in the revised definitions of distance functions and averaging operations. An example of a self-organizing map of attribute trees is shown in Fig. 7.12, indicating how the structure and the attributes of trees can be ordered in continuous fashion.

Figure 7.12: A self-organizing map of attribute trees.

References

[1] M. Peura. *Attribute trees as adaptive object models in image analysis.* D. Sc. thesis. Helsinki University of Technology, Laboratory of Computer and Information Science, February 2001.

[2] M. Peura and M. Syrjäsuo. Flexible Heuristic Matching of Attribute Trees. In *12th Scandinavian Conference on Image Analysis (SCIA'01)*, pages 207–213 Bergen, Norway, June 2001.

[3] M. Peura. Attribute trees in image analysis — heuristic matching and learning techniques. In *10th International Conference on Image Analysis and Processing, (ICIAP'99)*, Venice, Italy, September 1999. IEEE Computer Society.

Chapter 8

On-line recognition of handwritten characters

Vuokko Vuori, Matti Aksela, Ramūnas Girdziušas, Jorma Laaksonen, Erkki Oja

8.1 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, since its beginning in late 1997, 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. The project has earlier been a part of the TEKES technology programme Adaptive and Intelligent Systems Applications (AISA) and a subproject of the research project IMPRESS – Intelligent Methods for Processing and Exploration of Signal and Systems. The work is carried out in co-operation with Nokia Research Center.

We have performed series of experiments both with laboratory's large-scale computers and real hand-held PDA (personal digital assistant) devices. So far, we have performed genuine on-line user experiments with the PDA and the adaptive character recognizer by using a specially designed questionnaire program. The results have been promising and shown that users find the autonomous adaptation of the system comfortable. Currently, we are implementing our character recognizer to a new PDA with Linux operating system so that it could be used for textual input with all kind of applications. This way, we could experiment the adaptive recognition methods with more realistic tasks such as note making and e-mail composing.

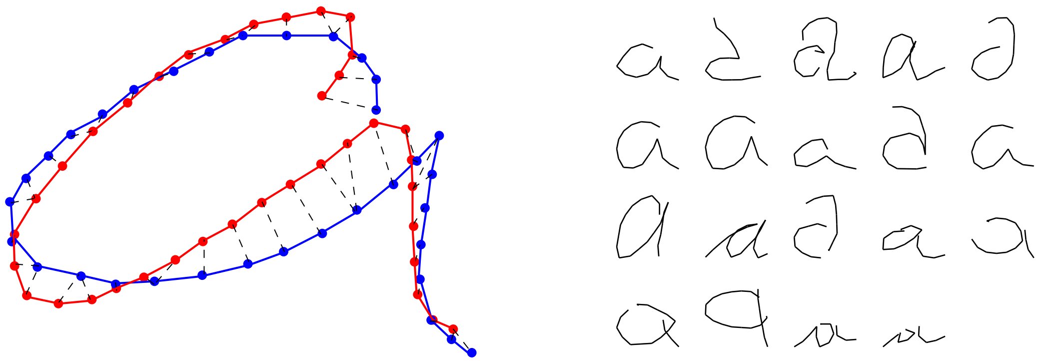


Figure 8.1: On the left: two handwritten characters 'a' are matched. On the right: a selection of different writing styles found for handwritten character 'a'.

8.2 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 perform reasonably well, quickly and accurately, with all kind of writers. The adaptation process of the recognition system 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. Ideally, the system would be able to learn a new style from a single sample. In addition, 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 deduced from the user's actions and responses to the recognition results by the system itself, instead of being specified by the user in some training mode. Naturally, the system should be robust against labelling errors of the 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.

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. In our work, we have applied similarity measures based on the Dynamic Time Warping (DTW) algorithm [1]. The computational complexity of the DTW algorithm depends quadratically on the average number of data points in the prototypes and character samples. Therefore, we 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. This way, the recognition time could be decreased by 76% while no new errors were introduced [2]. A faster classification can also be achieved by modifying the similarity measure by posing stricter constraints for the nonlinear matching of the data points performed by the DTW algorithm [3].

Another approach to speed up 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) [4]. These experiments showed that clusters of writing styles can be found but the writers cannot be realiably assigned to them on the basis of a small set of arbitrary character samples [5].

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 these three modes of adaptation are used together. If the first mode of adaptation is used alone, the recognition rate improves quickly but the size of the prototype set tends to grow considerably. The second mode of adaptation is also useful, the recognition rates improve and the size of the prototype set remains the same. However, it is not sufficient when used alone if the user's character samples and the prototypes are too different. For example in our work, character samples and prototypes are compared stroke-wise and a new prototype is needed if the number of strokes in the character matches with none of existing correct prototypes [6]. According to our experiments, only two new prototypes per class would be enough for adapting the recognition system to a writing style [7]. 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 [8].

References

[1] D. Sankoff and J. B. Kruskal. *Time warps, string edits, and macromolecules: the theory and practice of sequence comparison.* Addison-Wesley, 1983.

[2] Vuokko Vuori, Jorma Laaksonen, Erkki Oja, and Jari Kangas. Speeding up on-line recognition of handwritten characters by pruning the prototype set. In *Proceedings of 6th International Conference on Document Analysis and Recognition*, Seattle, Washington, USA, pages 501–505, September 2001.

[3] Vuokko Vuori, Matti Aksela, Jorma Laaksonen, Erkki Oja, and Jari Kangas. Adaptive character recognizer for a hand-held device: implementation and evaluation setup. In L. R. B. Schomaker and L. G. Vuurpijl, editors, *Proceedings of the Seventh International Workshop on Frontiers in Handwriting Recognition*, pages 13–22, Amsterdam, September 2000. Nijmegen: International Unipen Foundation. ISBN 90-76942-01-3.

[4] Teuvo Kohonen. *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences*. Springer-Verlag, 1997. Second Extended Edition.

[5] Vuokko Vuori and Erkki Oja. Analysis of different writing styles with the self-organizing map. In *Proceedings of the 7th International Conference on Neural Information Processing*, volume 2, pages 1243–1247, November 2000.

[6] Vuokko Vuori, Jorma Laaksonen, Erkki Oja, and Jari Kangas. Experiments with adaptation strategies for a prototype-based recognition system for isolated handwritten characters. *International Journal on Document Analysis and Recognition*, 3(3):150–159, March 2001.

[7] Vuokko Vuori, Jorma Laaksonen, Erkki Oja, and Jari Kangas. Controlling on-line adaptation of a prototype-based classifier for handwritten characters. In *Proceedings of the 15th International Conference on Pattern Recognition*, volume 2, pages 331–334, September 2000.

[8] 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.

8.3 Adaptive committee techniques

Matti Aksela, Jorma Laaksonen, Erkki Oja

Combining several different classifiers in a committee form is another way to reach for the best attainable recognition performance. Combining the results of several classifiers can improve performance because in the outputs of the individual classifiers the errors are not necessarily overlapping. Committee methods generally require more than one member classifier to recognize the input. In the near future, this will not be computationally too complex for even the smallest platforms for on-line handwritten character recognition, due to the continuous increase in available computational power.

The objective of a committee classifier is to combine the results of a set of member classifiers in a way that improves the overall performance. The two most important features of the member classifiers that affect the committee's performance are their individual error rates and the correlatedness of the errors. The more different the mistakes made by the classifiers, the more beneficial the combination of the classifiers can be.

We have experimented with several adaptive committee structures. Two of the most effective methods examined so far have been the Dynamically Expanding Context (DEC) [1] and the novel Class-Confidence Critic Combining (CCCC) scheme. The DEC algorithm was originally developed to create transformation rules that would correct typical coarticulation effects in phonemic speech recognition. The main idea behind the approach is to determine just a sufficient amount of context for each individual segment so that all conflicts in the set of training samples will be resolved. The DEC principle has been slightly modified to suit the setting of isolated handwritten character recognition [2]. In the DEC committee, the classifiers are initialized and ranked in the order of decreasing performance. The primary and the second-ranking results of every member classifier 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 performed by that particular member classifier. The confidence value is obtained through constructing and updating distance distributions of each class in every critic. The distance distributions model the distance from the input sample to its nearest prototype. 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 [3].

References

[1] T. Kohonen. Dynamically expanding context. *Journal of Intelligent Systems*, 1(1):79–95, 1987.

[2] M. Aksela. Handwritten character recognition: A palm-top implementation and adaptive committee experiments. Master's thesis, Helsinki University of Technology, 2000.

[3] M. Aksela, J. Laaksonen, E. Oja, and J. Kangas. Application of adaptive committee classifiers in on-line character recognition. In *Proceedings of International Conference on Advances in Pattern Recognition*, pages 270–279, 2001.

8.4 Discriminative classifiers

Ramūnas Girdziušas, Jorma Laaksonen, Erkki Oja

Hidden Markov Model (HMM) classifiers [1] have been considered as an alternative to the DTW distance classifiers in order to produce recognizers with uncorrelated recognition errors. HMMs are probabilistic generative models whose parameters are estimated by optimizing a cost criterion. Usually HMM classifiers provide a maximum log-likelihood (ML) estimate C_{ML} for sequences of a particular class. Alternatively, an HMM can be trained to discriminate between the classes by maximizing the mutual information (MMI) measure C_{MMI} between the observed sequences and their models. The ML approach is numerically efficient, but often results in classifiers with over-fitted parameters. On the other hand, the MMI alternative allows attaining lower recognition error rates, but it requires computational resources that are not available when one considers palm-top device implementations. Both approaches are well justified on statistical grounds, but they yield sub-optimal recognition machines when confronted with practical requirements.

An important observation is that the criteria C_{ML} and C_{MMI} are functions of the HMM scaling factors [1]. We have shown [2] that under certain assumptions, these quantities can be regarded as the average mutual information between the generated sequence and the model's state at a particular time instant. The discriminant functions of the classifiers obtained by optimizing C_{ML} and C_{MMI} have been replaced by a kernel function expansion in the scaling factors, designed akin to the support vector machine (SVM) methodology [3]. The classifier functions in such a way that the character sequence is first mapped into very high-dimensional feature space. In that space, each dimension represents the mutual information between the generated sequence and the state occurrence at a fixed time instant, averaged over the states of the prototype model. Linear binary separating hyperplanes are then built to maximize the class-separating margin.

The created HMM-SVM classifier was shown to have certain advantages over the traditional HMM approaches. The method decreases the error rate by 12 %-units compared with the ML-HMM classifier, to the level of 25%. The developed classifier attains the same error rates as the SVM classifier working on character bitmaps. However, it requires less than 300 character prototypes for writer-independent task rather than roughly 6000 support vectors extracted from the character bitmaps. Compared to the MMI approach, the HMM-SVM does not require gradient ascent optimization in the high-dimensional parameter space. The within-kernel parameters are determined by the ML design and the coefficients in the kernel function expansion are determined by the quadratic programming algorithm. Its dual formulation has computational complexity independent of the number of prototype models and the length of the character sequences. Also, the training phase does not have the local minima problem inherent in the MMI approach.

References

[1] L. Rabiner and B. H. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, 1993.

[2] R. Girdziušas. Discriminative on-line recognition of isolated handwritten characters. Master's thesis, Helsinki University of Technology, November 2001.

[3] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Statistics for Engineering and Information Science. Springer, 2000.

Chapter 9

Self-organizing map

Teuvo Kohonen, Merja Oja, Samuel Kaski, Panu Somervuo

9.1 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 4000 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.

9.2 4465 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 4465 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]). Additions to the list and error reports are most welcome; please send any correspondence to the email address `biblio@mail.cis.hut.fi`.

Evolution of SOM applications. We studied how SOM research within certain application areas has evolved during the years. A set of 13 topical categories was selected by combining classes used by the INSPEC (tm) database. The number of articles in those categories was then plotted as a function of the year of publication. The plots only contain the articles available in the INSPEC collection.

The results shown in Figure 9.1 reveal, for instance, that speech recognition was a very popular topic already in the beginning of 90's, while there still are applications in that area. Some other disciplines, including "information science and documentation" and "business and administration," are still gaining popularity.

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 [2] 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] 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.icsi.berkeley.edu/∼jagota/NCS/: Vol 1, pp. 102–350.

[2] 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.

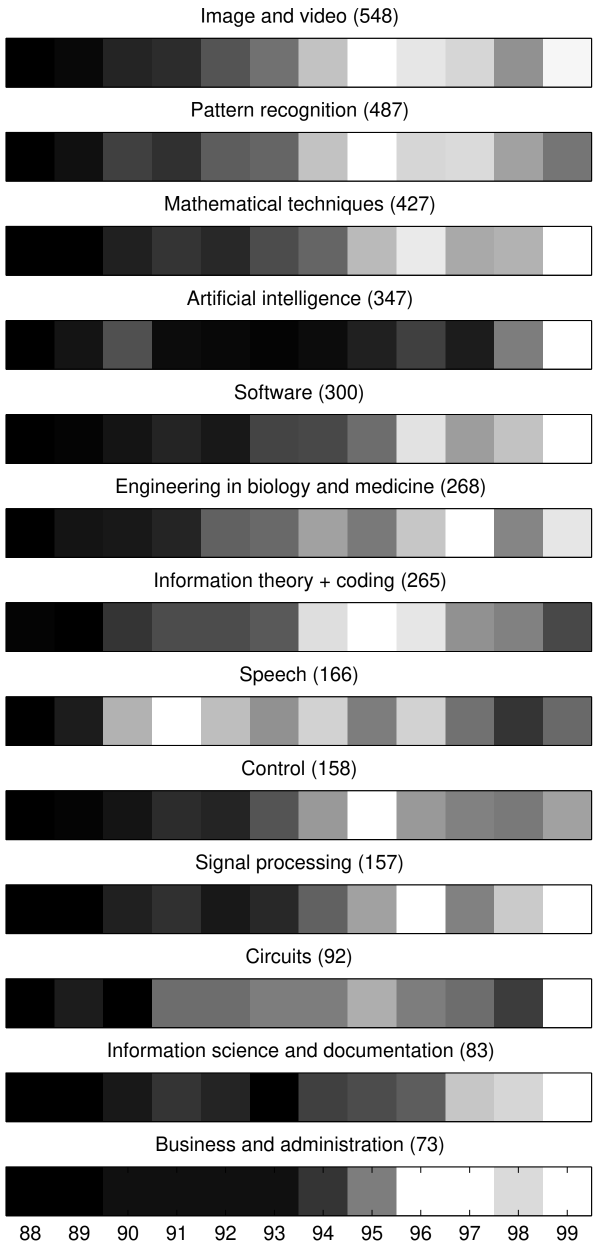


Figure 9.1: Evolution of SOM applications. Each horizontal bar shows the number of published SOM works within a certain topic category, as a function of the publication year shown in the bottom. The total number of works in the category is shown in parentheses. White: Largest number within the category, black: relatively small number of articles.

9.3 Clustering and visualization of large protein sequence databases by means of an extension of the self-organizing map

Panu Somervuo and Teuvo Kohonen

The amount of DNA sequences, protein sequences, and molecule structures studied and reported, e.g., in the Internet is already overwhelming. One should develop better tools for the analysis of the existing databases. Among the new challenges one may mention finding the hidden relations between the data items, revealing structures from large databases, and representing the results to the human in a comprehensible way. The classification and clustering of the sequences may reveal new unknown connections between them. The visualization of large data sets in a compact way may give insights into the data and lead to the development of new ideas and theories.

The Self-Organizing Map of symbol strings [2] was used in this work [4] for the clustering of all the 77,977 protein sequences of the SWISS-PROT database, release 37 [1]. In this method, unlike in some previous ones, the data sequences are not converted into histogram vectors in order to perform the clustering. Instead, a collection of true representative model sequences that approximate the contents of the database in a compact way is found automatically, based on the concept of the generalized median of symbol strings, after the user has defined any proper similarity measure for the sequences. The FASTA method [3] was used in this work. The benefits of the SOM and also those of its extension are fast computation, approximate representation of the large database by means of a much smaller, fixed number of model sequences, and an easy interpretation of the clustering by means of visualization. The complete sequence database is mapped onto a two-dimensional graphic SOM display, and clusters of similar sequences are then found and made visible by indicating the degree of similarity of the adjacent model sequences by shades of gray. The geometrically organized picture makes it possible to illustrate the relationships of a large amount of sequences at a glance, see Fig. 9.2.

Besides the protein sequences, we have applied the extension of the SOM also to the clustering of the protein molecules based on their three-dimensional molecule structures, see Fig. 9.3.

References

[1] A. Bairoch and R. Apweiler. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. *Nucleic Acids Res.*, 27:49-54, 1999.

[2] T. Kohonen. Self-organizing maps of symbol strings. Report A 42, Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, Finland, 1996.

[3] W. Pearson and D. Lipman. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA*, 85:2444-2448, 1988.

[4] P. Somervuo and T. Kohonen. Clustering and Visualization of Large Protein Sequence Databases by Means of an Extension of the Self-Organizing Map. 3rd International Conference on Discovery Science, Kyoto, Japan, Dec. 4-6, 2000, *Lecture Notes in Artificial Intelligence* 1967, pages 76–85, 2000.

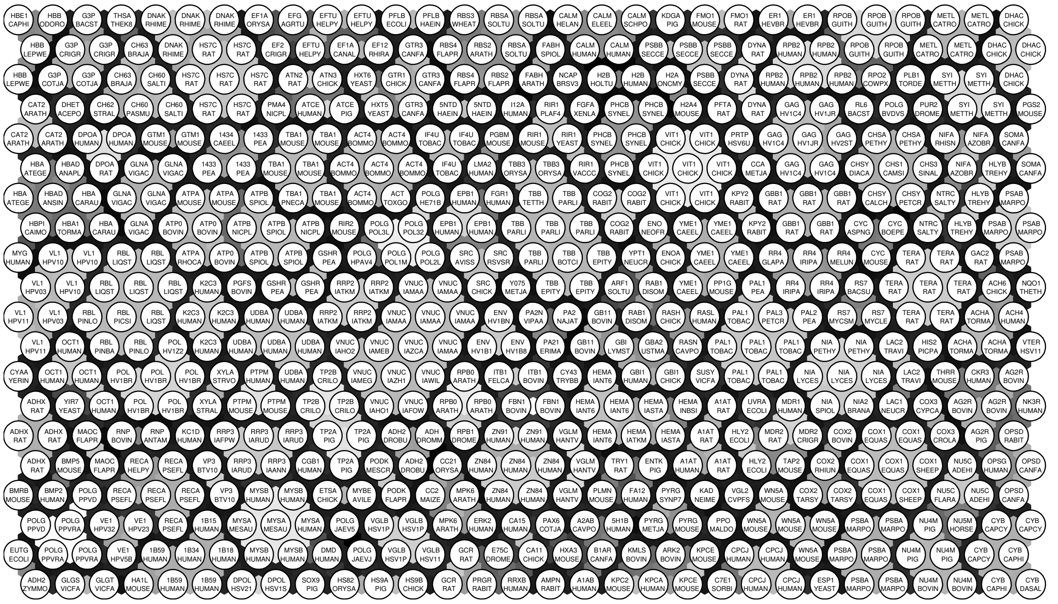


Figure 9.2: A 30-by-20-unit hexagonal SOM grid. The SOM was constructed using all the 77,977 protein sequences of the SWISS-PROT release 37. Each node contains a prototype sequence and a list of data sequences. The labels on the map nodes are the SWISS-PROT identifiers of the prototype sequences. The upper label in each map node is the mnemonic of the protein name and the lower label is the mnemonic of the species name. The similarities of the neighboring prototype sequences on the map are indicated by shades of gray. The light shades indicate a high degree of similarity, and the dark shades a low degree of similarity, respectively. Light areas on the map reveal large clusters of similar sequences.

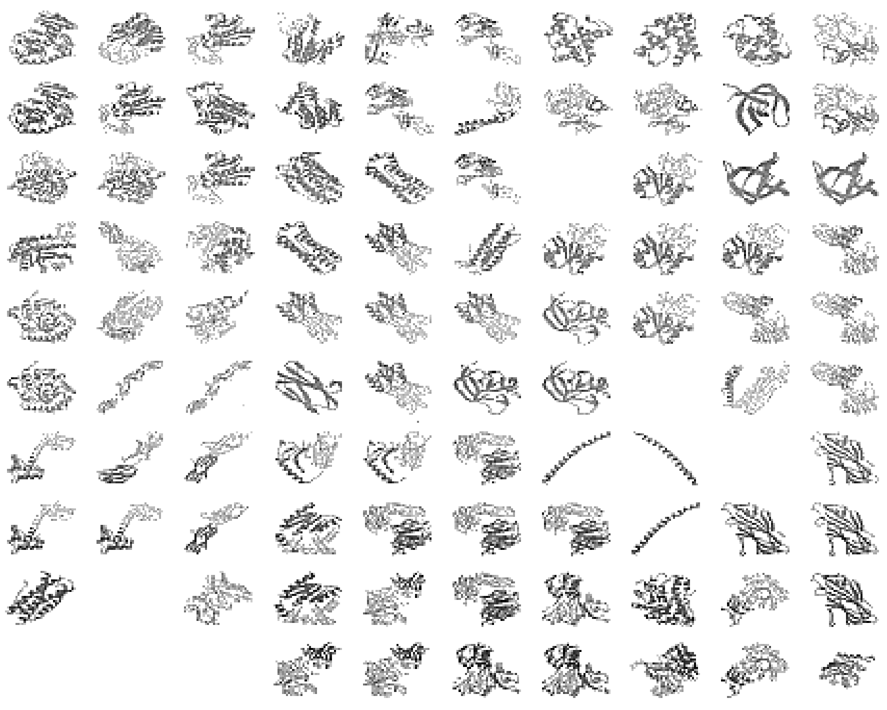


Figure 9.3: The SOM of 3D protein molecules.

Chapter 10

Self-organization of very large document collections

Teuvo Kohonen, Samuel Kaski, Krista Lagus, Jarkko Salojärvi, Jukka Honkela, Vesa Paatero, Antti Saarela

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 [6]. The SOM, applied to organizing very large document collections, can aid in all the three tasks.

Statistical models of documents

In the vast majority of SOM applications, the input data constitute high-dimensional real *feature vectors*. In the SOMs that form similarity graphs of *text documents*, the feature sets that describe collections of words in the documents should be expressible as real vectors, too. The feature sets can simply be weighted histograms of the words, but usually some dimensionality reduction of the histograms is carried out.

The primitive vector space model

In the basic *vector space model* [1] the stored documents are represented as real vectors in which each component corresponds to the frequency of occurrence of a particular word in the document: the model or document vector can be viewed as a weighted word histogram. For the weighting of a word according to its importance one can use the Shannon entropy over document classes, or the inverse of the number of the documents in which the word occurs ("inverse document frequency"). The main problem of the vector space model is the large vocabulary in any sizable collection of free-text documents, which means a vast dimensionality of the model vectors.

Latent semantic indexing (LSI)

In an attempt to reduce the dimensionality of the document vectors without essentially losing information contained in the full vocabulary, one often first forms a matrix in which each column corresponds to the word histogram of a document, and there is one column for each document. After that the space spanned by the column vectors is decomposed into an ordered set of factors by a matrix-computation method called the singular-value decomposition (SVD). The decomposition has the property that the last factors have minimal influence on the matrix. When they are omitted the document vector formed from the histogram of the remaining factors has then a much smaller dimensionality while as much as possible is retained of the original histograms. This method is called *latent semantic indexing (LSI)* [2].

Randomly projected histograms

We have shown that the dimensionality of the document vectors can be reduced radically by a much simpler and computationally lighter method than the LSI, the random projection method [3], without essentially losing the power of discrimination between the documents. Consider the original document vector (weighted histogram) $\mathbf{n}_i \in \mathbb{R}^n$ and a rectangular random matrix $\mathbf{R}$, the elements in each column of which are assumed to be normally distributed. Let us form the document vectors as the *projections* $\mathbf{x}_i \in \mathbb{R}^m$, where $m \ll n$:

$$\mathbf{x}_i = \mathbf{R}\mathbf{n}_i \; . \tag{10.1}$$

It can be shown that for large m, the similarity relations between arbitrary pairs of projection vectors $(\mathbf{x}_i, \mathbf{x}_j)$ are very good approximations of the corresponding relations between the original document vectors $(\mathbf{n}_i, \mathbf{n}_j)$, and the computing load of the projections is reasonable. The variance of the error is inversely proportional to m, and in our experiments values of m exeeding 100 have yielded practically as good results as the original vectors (see [3] for theoretical considerations and [5] for comparison results). On the other hand, with the radically decreased dimensionality of the document vectors, the time needed to classify a document is radically decreased.

Construction of the random projections by pointers

There exists a special method for forming projections that gives as good results as the random projections discussed above but is computationally much more efficient for sparse input vectors. This method is used in our latest experiments, described e.g. in [5].

Construction of the document map

A brief overview of the computing phases is given in the following.

Preprocessing. From the raw text, nontextual and otherwise nonrelevant information (punctuation marks, articles and other stopwords, message headers, URLs, email addresses, signatures, images, and numbers) was removed. The most common words, and words occuring rarely (e.g., less than 50 times in the corpus) were also discarded.

The variation caused by different word forms and inflections has been reduced in our later experiments by using a *stemmer* for both Finnish and English text.

Formation of statistical models. To reduce the dimensionality of the models, we have used both randomly projected word category histograms (see, e.g., [5]) and randomly projected word histograms (or their shortcut computation by pointers), weighted by the Shannon entropy or the “inverse document frequency” (idf).

Formation of the document map. The document maps were formed automatically by the SOM algorithm, for which the statistical models of documents were used as input. The size of the SOM was determined so that on the average 7 articles were mapped onto each grid point; this figure was mainly determined for the convenience of browsing.

The speed of computation, especially of large SOMs can be increased by several methods. In our latest experiments we have used the Batch Map algorithm [5], and furthermore, the winner search was accelerated in several ways. The search was started in the neighborhood of corresponding winners at the previous cycle of iteration. The distance computation, or in this case actually computation of inner products, can be speeded up even further by neglecting the components having the value zero; a large proportion of the components are zeroes since the input vectors are still sparse after the dimensionality has been reduced by the random projection with pointers.

The size (number of grid nodes) of the maps was increased stepwise by estimating a larger map based on a smaller one. After each increase the winners for all data vectors can be found quickly based on the estimation formula used for increasing the map. If necessary, the winner search can be speeded up even further by restricting the search to the area of the larger map that corresponds to the neighborhood of the winner on the smaller map.

For very large maps it may be necessary to reduce the amount of memory needed for storing the maps. We have represented the model vectors with reduced accuracy and used probabilistic updating to maintain numerical accuracy in adapting the model vectors.

To evaluate the performance of the speedup methods in constructing document maps we compared document maps computed using the shortcut methods with maps computed by the traditional SOM algorithm. As can be seen from Table 10.1, the quality of the resulting document maps is comparable, but the time needed for the shortcut methods is only about one tenth of that of the traditional algorithm.

Table 10.1: Comparison of the shortcut methods with the traditional SOM algorithm. The figures are averages from five test runs with different random matrices used in the encoding of the documents, and the error margins are standard deviations. The data was 13,742 abstracts from 21 patent classes, and the final maps consisted of 1,344 units. The time has been measured with a SGI O2000 computer without parallelization of any programs.

	Classification accuracy (%)	Quantization error	Time (s)
Traditional SOM	58.2 ± 0.2	0.799 ± 0.001	2550 ± 40
Shortcut methods	58.0 ± 0.2	0.798 ± 0.002	241 ± 3.5

For very large maps the difference in the computation times is even more marked than in Table 10.1, but can only be deduced from the computational complexities given in Table 10.2.

User interface. The 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 [8], has been

Table 10.2: Computational complexity of the methods. Here N denotes the number of data samples, M the number of map units in the small map, and d the dimensionality of the input vectors. It has been assumed that the number of map units in the final map is chosen to be proportional to the number of data samples.

	Computational complexity
Traditional SOM	$\mathcal{O}(dN^2)$
Shortcut methods	$\mathcal{O}(dM^2) + \mathcal{O}(dN) + \mathcal{O}(N^2)$

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.

If the document map is very large, the comparison between the document vector and all the model vectors is time-consuming. It is, however, possible to make rapid approximations by restricting the comparisons in such subspaces of the original space that best represent the (local) organization of the map.

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.

Example: The largest published map

For the largest WEBSOM map made so far we selected a data base of 6,840,568 patent abstracts available in electronic form and written in English. These patents were granted by the U.S., European, and Japan patent offices and stored in two databases: the "First Page" database (1970-1997), and "Patent Abstracts of Japan" (1976-1997). The average length of the texts was 132 words. The size of the finally accepted vocabulary was 43,222 words. The words were weighted by the Shannon entropy computed from the distribution of the words into 21 subsections from the IPC6 patent classification.

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 accuracy of classifying a document into one of the 21 subsections was about 64 per cent. It must be noted that the categories are in fact overlapping: many patents belong to several different categories.

Figure 10.1 exemplifies a case of content-addressable search. The map and the collection can be explored using a WWW-browser. Further examples of document maps can be found at `//websom.hut.fi/websom/`.

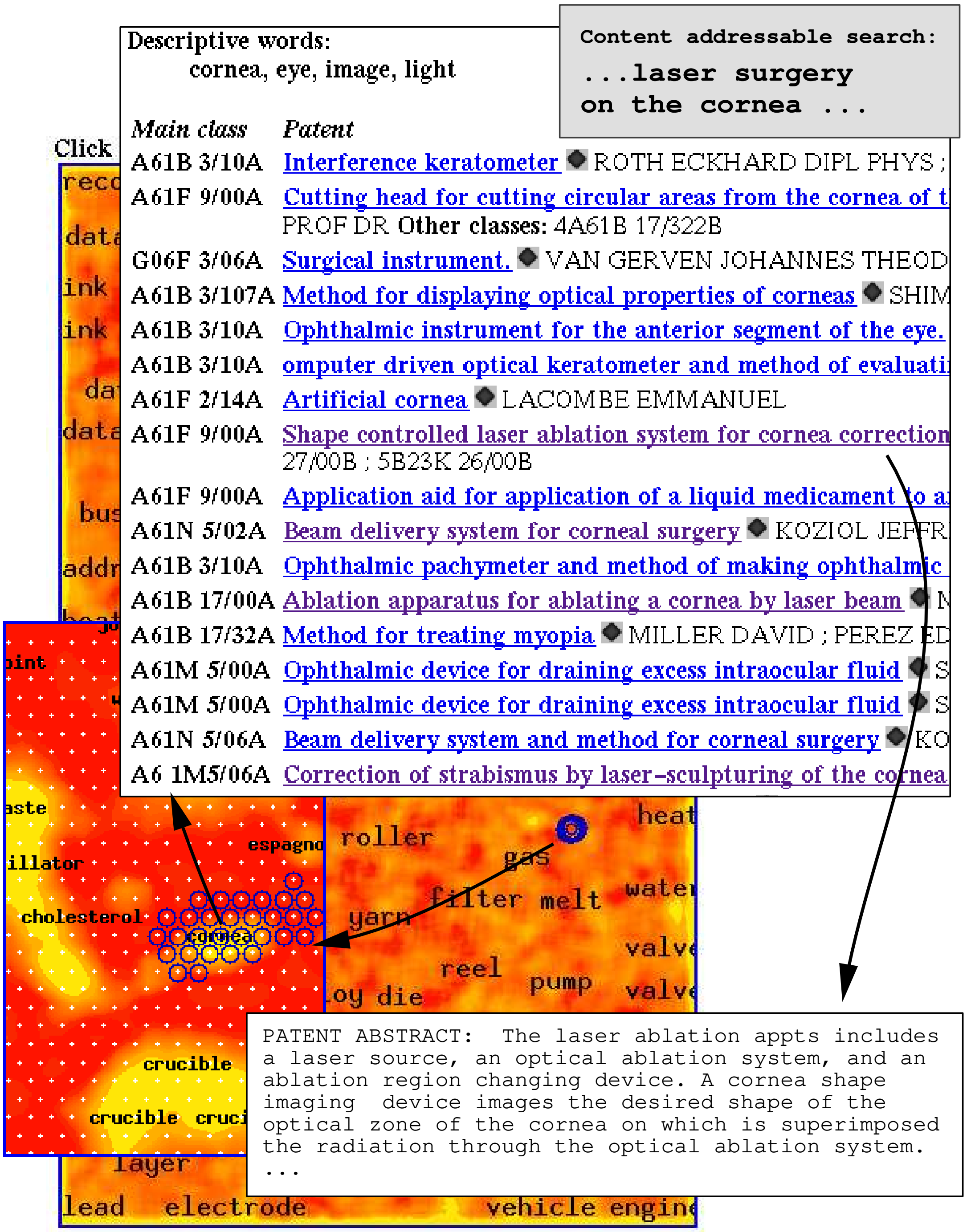


Figure 10.1: Content addressable search was utilized on a WEBSOM document map of nearly 7 million patent abstracts. The document map is depicted in the background, and the shades of gray correspond with document clusters. The search query was “laser surgery on the cornea” and the best-matching map locations are marked with circles. Zooming on the area reveals a small cluster of map units that contains patent abstracts mostly about the cornea of the eye, and of surgical operations on it. The words written on the document map have been selected automatically using the method described in [8].

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 [7] and in the following, the document maps can also be applied for searching without the benefit of the visual interface. This allows comparison to systems that do not offer any visualization or exploration capabilities.

The search algorithm

Consider the typical search task where one wishes to locate the n best documents corresponding to a query, where n is a small number, say 50, and the documents are to be ordered based on goodness of match. The following algorithm is proposed for performing such searches using a document map:

1. Off-line organizatory phase: Cluster the documents on a WEBSOM document map with M map units. Find the best-matching unit on the map for each document. Represent the set of documents within a map unit by the model vector of that unit.
2. On-line search phase: For a given query,
   (a) *Pre-select:* Compare the query vector with the model vectors to select the best map units (based on minimum inner product distance from the query vector), and collect documents from those units until a total of K documents ($K >= n$) are obtained.
   (b) *Refine:* perform an exhaustive search among the K prospective documents and return the n best ones in the order of goodness. In the straightforward case the exhaustive search amounts to comparing the document vectors to the query vector, and measuring similarity by inner products. Alternatively, a more detailed search could be carried out.

The hierarchical search is computationally efficient: The full index of the document collection need not be held in main memory, nor is there a need to search through all the documents.

Experiment on text retrieval

The usefulness of the approach was studied on the well-known CISI reference collection for text retrieval. The CISI data set consists of 1460 documents and 76 queries. The collection lists for each query all the documents that are considered relevant to the query (41 on the average), as judged by a human. The average lengths of the documents and the queries were 115 and 51 words, respectively. The CISI collection is generally considered difficult for retrieval systems.

Creation of the document map. The documents were organized on a WEBSOM map of 150 units. The preprocessing was carried out as described above. The size of the vocabulary was 2136 words, small enough that no random projection was applied in document encoding.

Selecting a value for the K parameter. The set of queries was split randomly to two parts. The first half was used to select a value for the K parameter (the number of documents chosen at the pre-selection phase) and the second half for evaluating the results.

Measuring accuracy of retrieval. The accuracy was measured using *non-interpolated average precision over all relevant documents* (AP) [9]. AP is calculated for a single query by recording the precision each time a relevant document is retrieved, and taking an average of these (the precision of non-retrieved documents is taken to be zero). The results are further averaged over all queries. The measure not only takes into account the *number* of relevant documents retrieved, but also their *order* in the result set, rewarding systems for ranking relevant documents high in the returned lists of results.

Results. The results of a comparison between the proposed two-stage method, vector space model, and LSI are shown in Figure 10.2. The improvement obtained using docu-

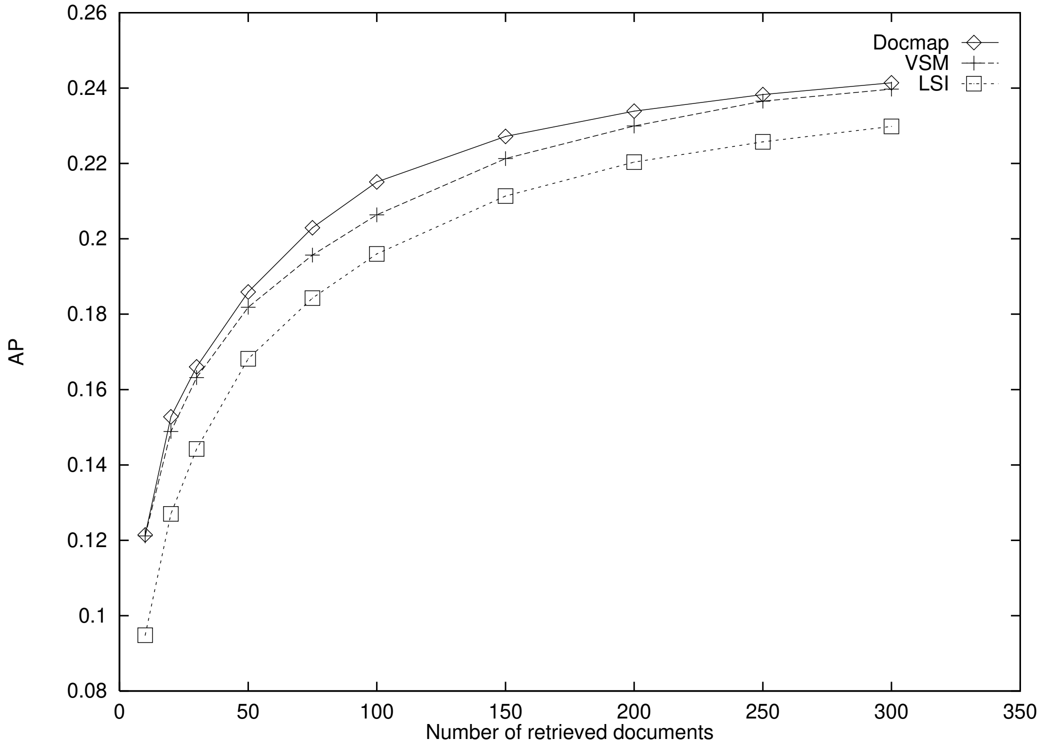


Figure 10.2: Comparison of methods. The average precision (AP) was calculated using the suggested document map method (Docmap), the vector space model (VSM) and the Latent Semantic Indexing (LSI) with various numbers of retrieved documents. The results are averages over the queries in the second half of the test set.

ment maps when compared to the vector space model was found to be statistically significant (e.g. for $n = 100$, in paired t-test p=0.00017). The difference between Docmap and LSI, although larger on the average, was not statistically significant due to considerably larger variation.

In effect, the model vectors representing clusters of similar documents obtained by the WEBSOM method are smoothed (i.e., generalized) representations of possible (sub-)topics of interest. The two-stage search thus first selects a small number of relevant (sub-)topics, then among those, picks the most relevant documents. Thus, the pre-selection phase serves to discard documents that have only accidental resemblance to the query, since they belong to a sub-topic that does not as a whole seem relevant enough for the query.

Conclusions

We have demonstrated that it is possible to scale up the SOMs in order to tackle very large-scale problems. Additionally, it has transpired that the encoding of documents for their statistical identification can be performed much more effectively than believed a few years ago [2]. In particular, the various random-projection methods are as accurate in practice as the ideal theoretical vector space method, but much faster to compute than the eigenvalue methods (e.g., LSI) that have been used extensively to solve the problem of large dimensionality.

Finally it ought to be emphasized that the order that ensues in the WEBSOM may not represent any taxonomy of the articles and does not serve as a basis for any automatic indexing of the documents. 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 the search task.

References

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

[2] Deerwester, S., Dumais, S., Furnas, G., Landauer, K. Indexing by latent semantic analysis. *J Am Soc Inform Sci*, 1990; 41:391–407.

[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] Kohonen, T. New developments of learning vector quantization and the self-organizing map. *In SYNAPSE´92, Symposium on Neural Networks; Alliances and Perspectives in Senri 1992*, Osaka, Japan, June 24-26, 1992.

[5] 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.

[6] Lagus, K. Text Mining with the WEBSOM. *Acta Polytechnica Scandinavica, Mathematics and Computing Series No. 110*, 54 pp. December 2000. D.Sc(Tech) Thesis, Helsinki University of Technology, Finland.

[7] Lagus, K. Text retrieval using self-organized document maps. *Neural Processing Letters*. Accepted for publication. 2002.

[8] 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.

[9] Voorhees, E. M., and Harman, D. K. Appendix: Evaluation techniques and measures. In *Proceedings of The Eighth Text REtrieval Conference (TREC 8)*. NIST, 2000.

Chapter 11

Bioinformatics

Samuel Kaski, Janne Nikkilä, Merja Oja

11.1 Introduction

Bioinformatics refers to the study of biological data using methods from mathematics, statistics, and computer science. In particular, functional genomics experiments produce massive amounts of high-dimensional data that need to be analyzed and understood [7, 8]. Recently developed methods such as oligonucleotide arrays and DNA chips (microarrays) can be used to measure the activity or expression level of each gene in a genome as a function of time in various experimental settings.

The need for data driven methods for generating hypotheses about gene function is already acknowledged in the functional genomics community (see for example [11]). Exploratory methods for information visualization, developed in the Neural Networks Research Centre, would provide precisely the required tools in this first stage of analysis. Methods based on the learning metrics principle (Section 12), on their part, are needed to focus the analyses on the important parts of the very high-dimensional and noisy data, by finding the dependencies between the gene expression data, known functional classes of genes, and other kinds of biological databases including the gene sequences and properties of the corresponding proteins.

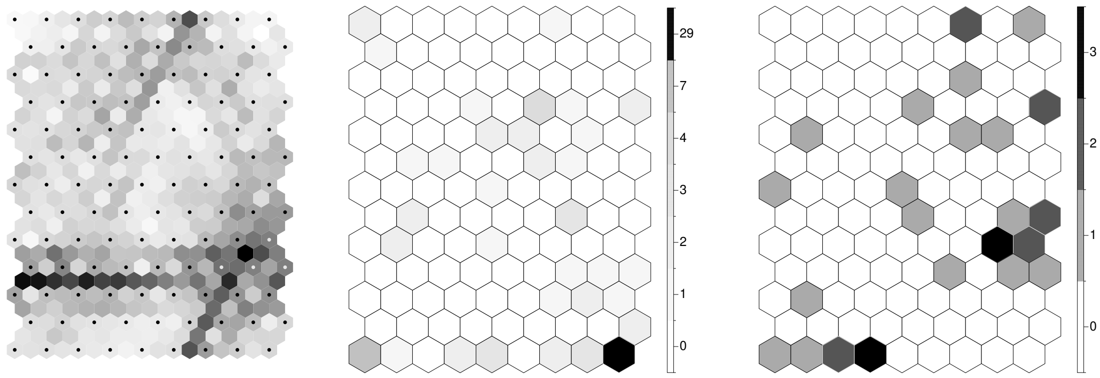
The project is carried out in collaboration with experts of the biological problem, and with the other bioinformatics group of the laboratory led by prof. Mannila and prof. Hollmén. The studies were started in late 2000 in collaboration with a group of the University of Kuopio led by Prof. Eero Castrén.

11.2 SOM-based exploration of gene function

We started by exploratory data analysis of gene expression [3, 4, 9]. The Self-Organizing Map (SOM) is particularly useful in this first stage of data analysis. The SOM constructs a nonlinear projection of the data to a map display which can be used for visualizing of similarity relationships and cluster structures, with methods developed in the Neural Networks Research Centre. Such a combination of non-parametric clustering and visualization distinguishes the SOM from the many clustering methods commonly applied to gene data.

The same SOM display can be used for visualizing the relationships between data sets, such as the gene expression and the functional classes of the genes below.

The genes of the yeast *Saccharomyces Cerevisiae* were first clustered based on their expression in a set of different conditions and treatments such as the diauxic shift and a heat shock (in a public-domain data set). Visualizations of the cluster structure and the relationships of the clusters to the functional classes of the genes were constructed, and the clusters were interpreted in terms of the functional classes of the genes and their activity in the treatments. Figure 11.1 shows, for instance, that most genes related to cytoplasmic degradation form a cluster and hence are expressed similarly in the set of treaments. We demonstrated how to use SOMs in the exploratory task and proposed new hypotheses on the relationships between some functional classes.



a **b** **c**

Figure 11.1: **a** SOM-based visualization of the cluster structure in yeast gene expression data (from [1]). Light shades: clusters; dark shades: sparser areas or gaps in between clusters. The dots denote map units. Note that in this display there is a hexagon in between each pair of map units, whereas on the **b** and **c** figures only the map units themselves are shown. The same SOM is shown in all figures. **b**: Distribution of the functional class 'cytoplasmic degradation' (87 genes) on the SOM. **c**: Distribution of the functional class 'sugar and carbohydrate transporters' (32 genes).

One of the ultimate goals in functional genetics is understanding the regulatory pathways of the genes. This is the key to understanding the dependencies between the genes and to controlling the processes within the cell. The pathways have been studied by "knocking out" one gene at a time by mutations and inspecting the effects on gene expression. If the mutated gene is a vital part of some pathway, then the whole pathway will be blocked. We would thus expect to see all the critical genes in one pathway cluster together because they cause similar effects for the expressions of all the other genes in the cell.

We analyzed the similarities of mutated yeast strains with Self-Organizing Maps [9].

The clusters that had earlier been found by hierarchical clustering [2] were found by the SOM as well, verifying the viability of the method (Fig. 11.2). We were additionally able to propose some new groupings for the mutated yeast strains.

The conclusion from these first studies is that the SOM is a valuable addition to the toolbox of bioinformaticians.

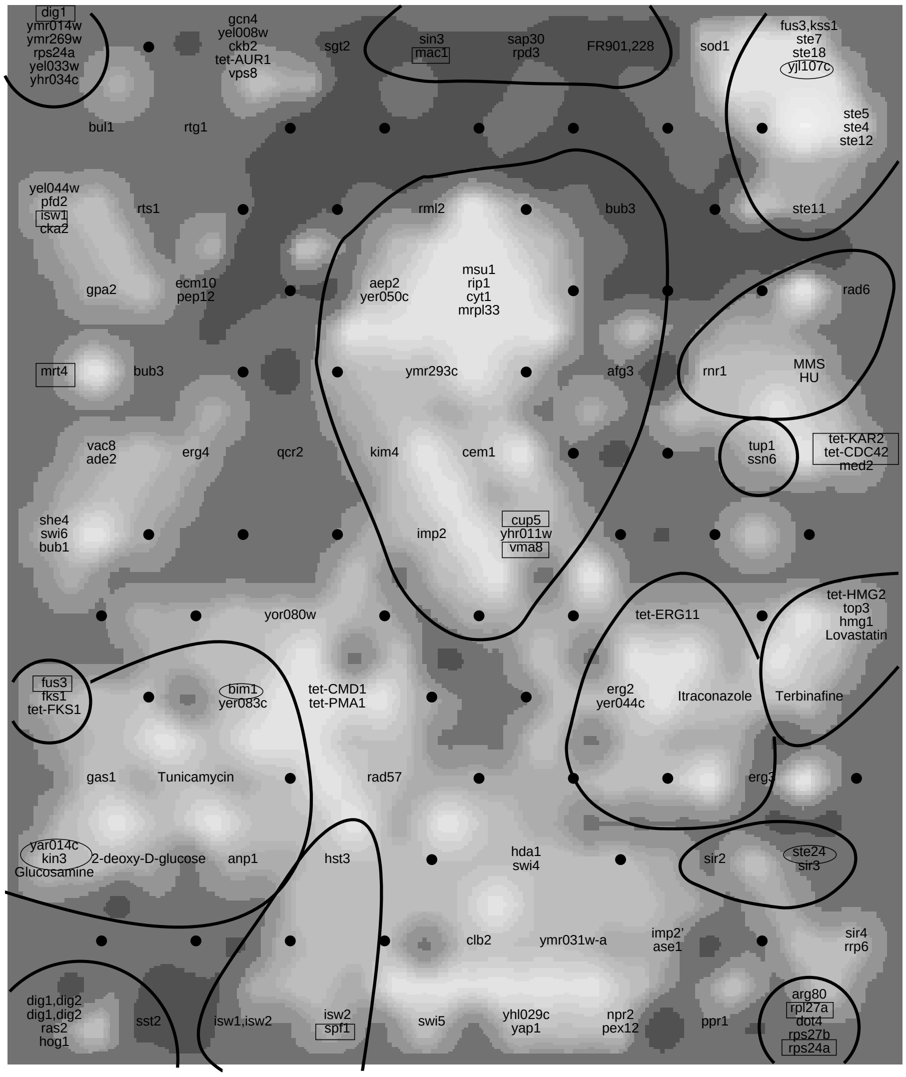


Figure 11.2: The Self-Organizing Map of the mutated yeast strains. Light shades denote dense areas in the expression space and dark sparse areas. The strains located near each other on the map are also nearby in the expression space. The manually drawn lines group the clusters derived earlier by hierarchical clustering [2] (small circles and boxes denote exceptions). As can be seen, approximately the same clusters can also be found based on this SOM-display. The SOM display additionally visualizes the similarities of the *clusters* and the data *between* them.

11.3 Discriminative clustering of genes

Proper selection of the metric of the gene expression space is a well-acknowledged problem since the data is high-dimensional, noisy, and contains uninteresting variation due to the several biological processes going on simultaneously within the cells. Our solution is to derive the metric from other biological data sets such as the functional classification (for details see Section 12).

More generally, there often exist several datasets derived from proteomics, gene expression, and genetic sequences, that could be combined to yield a more accurate picture of cell function.

Discriminative clustering (DC) [10] is a principled way to derive the metric used in clustering from combine auxiliary information. The first preliminary results of this approach applied to gene expression data were presented in [5, 10]. The DC was able to form clusters in the gene expression space that were more homogeneous with respect to the distribution of functional classes than the other methods. More detailed biological analysis and interpretation of these results is in progress.

The learning metrics principle can be applied to Self-Organizing Maps as well [6]. An application to gene expression data is in progress.

References

[1] 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, USA*, 95:14863–14868, 1998.

[2] Timothy R. Hughes, Matthew J. Marton, Allan R. Jones, et al. Functional discovery via a compendium of expression profiles. *Cell*, 102:109–126, 2000.

[3] Samuel Kaski, Janne Nikkilä, Petri Törönen, Eero Castren, and Garry Wong. Analysis and visualization of gene expression data using self-organizing maps. In *Proceedings of NSIP-01, IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing.* 2001.

[4] Samuel Kaski and Janne Sinkkonen. A topography-preserving latent variable model with learning metrics. In Nigel Allinson, Hujun Yin, Lesley Allinson, and Jon Slack, editors, *Advances in Self-Organizing Maps*, pages 224–229. Springer, London, 2001.

[5] Samuel Kaski, Janne Sinkkonen, and Janne Nikkilä. Clustering gene expression data by mutual information with gene function. In Georg Dorffner, Horst Bischof, and Kurt Hornik, editors, *Artificial Neural Networks—ICANN 2001*, pages 81–86. Springer, Berlin, 2001.

[6] 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.

[7] Nature Genetics Supplement. The chipping forecast. *Nature Genetics*, 21(1):1–60, 1999.

[8] Nature Insight. Functional genomics. *Nature*, 405:819–846, 2000.

[9] 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*. Kluwer, 2002. In press.

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

[11] Ognjenka Goga Vukmirovic and Shirley M. Tilghman. Exploring genome space. *Nature*, 405:820–822, 2000.

Chapter 12

Learning metrics

Samuel Kaski, Janne Sinkkonen, Jaakko Peltonen

12.1 Introduction

Visualization and clustering of multivariate data are usually based on mutual distances of samples, measured by heuristic means such as the Euclidean distance of vectors of extracted features. Our recently developed methods remove this arbitrariness by learning to measure important differences. The effect is equivalent to changing the metric of the data space. The laborious implicit supervision by manually tailored feature extraction will to a large extent be replaced by an automatically learned metric, while discoveries can still be made with unsupervised learning methods within the constraints set by the new metric.

It is assumed that variation of the data is important only to the extent it causes variation in *auxiliary data* c which is available paired to the primary data $\mathbf{x} \in \mathbb{R}^n$. Such auxiliary data are available at least in the settings where supervised learning methods, regression and classification, are usually applied. The difference here is that the goal is to model and understand the *primary data* and learn what is relevant there, whereas in supervised learning the sole purpose is to predict the auxiliary data.

In other words, important variation in $\mathbf{x}$ is supposed to be revealed 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.

$$d_L^2(\mathbf{x}, \mathbf{x} + d\mathbf{x}) \equiv D_{\mathrm{KL}}(p(c|\mathbf{x}) \| p(c|\mathbf{x} + d\mathbf{x})) = d\mathbf{x}^T \mathbf{J}(\mathbf{x}) d\mathbf{x} , \tag{12.1}$$

where $\mathbf{J}(\mathbf{x})$ is the Fisher information matrix. Bankruptcy risk is an example of auxiliary data that indicates importance in the analysis of the financial states of companies: the $c = 1$ if the company goes bankrupt and $c = 0$ if it stays alive. The Kullback-Leibler divergence is locally a metric, and locality can be relaxed by extending the metric. Proximity relations (i.e., loosely speaking, the topology) of the data space are preserved, but the arbitrariness of feature selection is removed by locally re-scaling the data space to make it reflect important variation in data.

The Fisher information matrix has earlier been used to construct metrics to spaces of probability models (see, e.g., [2]). The novelty here is that the information matrix is applied in the data space to construct a new metric there.

We call the idea of measuring distances in the data space by approximations of (12.1) the learning metrics principle. 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.

12.2 Learning metrics for Self-Organizing Maps

The principle

One way of approximating (12.1) is to first compute an estimator $\hat{p}(c|\mathbf{x})$ for the conditional density, and then compute the Fisher information matrix from the estimator. In practice the distance can be computed directly from

$$d_L^2(\mathbf{x}, \mathbf{x} + d\mathbf{x}) = E_{\hat{p}(c|\mathbf{x})} \left\{ \left(d\mathbf{x}^T \frac{\partial}{\partial \mathbf{x}} \log \hat{p}(c|\mathbf{x}) \right)^2 \right\} \; ; \tag{12.2}$$

the matrix $\mathbf{J}$ need not be formed explicitly.

We have applied this principle to the Self-Organizing Map (SOM) [4]. A SOM consists of a grid of computational units, with a model vector $\mathbf{m}_i$ associated to each unit i. After the SOM has been computed these model vectors follow the input data in an ordered fashion: model vectors of close-by units on the lattice remain close-by in the input space.

The SOM algorithm iterates two steps at discrete time steps t: winner selection and adaptation. In the new metric the index w of the unit closest to the current input $\mathbf{x}(t)$ is sought by

$$w(\mathbf{x}(t)) = \arg\min_i d_L^2(\mathbf{x}(t), \mathbf{m}_i(t)) \; . \tag{12.3}$$

Here the local distance approximation is also used for non-local distances, which has turned out to be reasonably accurate in many applications.

It can be shown that the adaptation step in the steepest descent direction of the new non-Euclidean metric, i.e. at the direction of the natural gradient [2], is equivalent to the familiar SOM learning rule,

$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) + h_{wi}(t)(\mathbf{x}(t) - \mathbf{m}_i(t)) \; . \tag{12.4}$$

Here $h_{wi}(t)$ is the neighbourhood function, a decreasing function of distance on the map lattice. The neighbourhood tightens slowly with the iterations.

In the case of the bankruptcy application described below, the computational complexity is approximately doubled compared to that of a SOM in the Euclidean metric.

A demonstration with a toy data is presented in Figure 12.1. In the learning metric the SOM finds the relevant horizontal dimensions of the data and does not waste resources on representing the irrelevant vertical dimension.

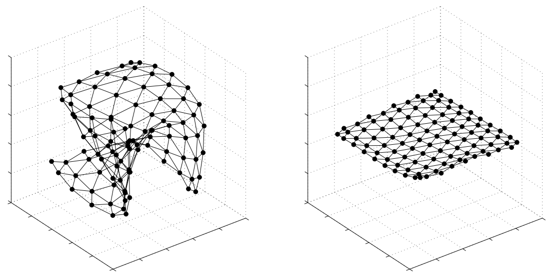


Figure 12.1: Projections of the model vectors of two SOMs representing the same data, in the Euclidean (left) and the learning metric (right). The primary data is evenly distributed within the unit cube. Only the horizontal dimension is relevant, indicated by the auxiliary data which is of constant distribution in the vertical dimension.

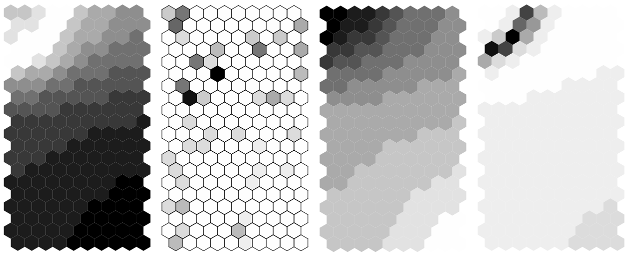


Figure 12.2: Sample visualizations of bankruptcy data with a SOM in the learning metrics. **a** Posterior bankruptcy probability, **b** empirical ratio of healthy to bankrupt companies, **c** distribution of the values of a profitability indicator, and **d** relevance of the profitability indicator. The hexagons correspond to SOM units and light shades denote high values. Each company is located in one hexagon.

Bankruptcy Analysis

Most quantitative studies of bankruptcy have aimed at prediction, mainly by classification and probability estimation based on financial statements. A complementary approach pursued earlier in our laboratory [5, 6] is to analyze and visualize the effects of corporate behaviour on the bankruptcy risk. We have applied learning metrics, which enables us to guide the visualizations by the most important indicator of bankruptcy analysis: whether the company goes (has gone) bankrupt or not. The SOM should then emphasize the most interesting features of the financial statements, i.e. those that contribute locally to the bankruptcies. The method automatically discards irrelevant features of input data, and therefore makes it possible to incorporate potential-looking financial indicators more freely that is possible with a completely unsupervised method.

In this case study, the data consisted of financial statements from Finnish companies. Several financial indicators were extracted from the statements, and used as primary data. The auxiliary variable indicated whether the company went bankrupt within 3 years of the statement.

Gaussian mixture-based estimates for the conditional density of bankruptcy were computed and used to build the metric for SOMs. As expected, the SOMs in the new metric were significantly more accurate in modeling bankruptcy risk than SOMs in the Euclidean metric. Visually, the learning metric displays were comparable or better than Euclidean displays. Sample visualizations are shown in Figure 12.2. The learning metric makes it possible to additionally visualize the local *relevance* of each variable, one at a time, in different locations of the map display (Fig. 12.2**d**).

12.3 Discriminative clustering

A general goal of clustering is to minimize within-cluster distortion or variation, and to maximize between-cluster variation. We have applied the learning metrics principle (12.1) to clustering by replacing the distortion measure within each local cluster by a kind of within-cluster Kullback-Leibler divergence. This causes the clusters to be internally as homogeneous as possible in terms of the distribution of the auxiliary data $p(c|\mathbf{x})$. The other side of the coin is that between-cluster differences in $p(c|\mathbf{x})$ are maximized; this is the reason for coining the method *discriminative clustering*.

The clusters are defined in terms of the primary data, which enables clustering of future samples without any auxiliary data. The locality of the clusters in the primary data space makes them useful for exploratory analysis.

In vector spaces

For vectorial data, we have parameterized the clusters as softened Voronoi regions, either in the Euclidean metric or, in some applications, the inner product metric of the unit hypersphere. The principle of stochastic approximation leads to very simple on-line learning algorithms which can be interpreted as modifications of the traditional Hebb rule.

The cost function of discriminative clustering is in general the distortion

$$E = \int \sum_j y_j(\mathbf{x}) D_{\mathrm{KL}}(p(c|\mathbf{x}) \| \boldsymbol{\psi}_j)\, p(\mathbf{x})\, d\mathbf{x} \,, \tag{12.5}$$

where the $y_j(\mathbf{x})$ denote the softened Voronoi regions ($\sum_j y_j(\mathbf{x}) = 1$), and $\boldsymbol{\psi}_j$ is a region-wise prototype for the conditional distribution $p(c|\mathbf{x})$. The cost function measures the inhomogeneity of the clusters w.r.t. $p(c|\mathbf{x})$, equivalent to a conditional likelihood for finite data and hard clusters.

Our cluster memberships have been of the form $y_j(\mathbf{x}) = Z^{-1}(\mathbf{x}) \exp(-d^2(\mathbf{m}_j, \mathbf{x})/\sigma^2)$, where $d(\mathbf{m}_j, \mathbf{x})$ is the distance between $\mathbf{x}$ and the *location prototype* $\mathbf{m}_j$, measured in the original metric of the data space. Then the on-line algorithm for minimizing the distortion takes a particularly simple form. For soft Euclidean Voronoi regions, an observation $\mathbf{x}$ with auxiliary information c causes simple updates to cluster locations:

$$\mathbf{m}_j := \mathbf{m}_j + \alpha(\mathbf{x} - \mathbf{m}_j) \log \frac{\psi_{li}}{\psi_{ji}} \,. \tag{12.6}$$

The clusters j and l are chosen randomly, proportionally to the activations $y_j(\mathbf{x})$.

So far the algorithm has been applied to clustering of text documents, yeast gene expressions, and financial statements of companies [8, 9]. In each of these applications, clusters with interesting variation are found by guiding the algorithm with suitable relevance-inducing auxiliary data.

It can be shown that the clustering maximizes mutual information between the auxiliary data and the clusters interpreted as a random variable. Other mutual information maximizing clustering methods have been proposed earlier [1, 10]. Our method combines a novel on-line algorithm with the induction of a proximity-preserving quantization to a continuous data space.

Of texts

We have extended the scope of the method to text documents, under the commonly used simplifying assumption that the documents are 'bags of words', finite-length samples from a multinomial distribution. All knowledge of the document's topical content is encoded into the parameters of the distribution, and hence we cluster the distributions. Discriminative clustering applied to texts is then called Discriminative Distributional Clustering.

The clusters are now defined as Voronoi regions in the parameter space of the distributions $\mathbf{q}$, with Kullback-Leibler divergence from the prototypes as the distance measure. The Voronoi regions are softened for computational reasons by $y_j(\mathbf{q}; \boldsymbol{\Theta}) = Z^{-1}(\mathbf{q}) \exp(-\kappa D_{KL}(\mathbf{q}, \boldsymbol{\theta}_j))$. Here $Z(\mathbf{q})$ normalizes the memberships of $\mathbf{q}$, $\boldsymbol{\theta}_j$ is the distributional prototype of cluster j, and $\boldsymbol{\Theta}$ denotes all the parameters of the membership functions.

The method has been applied to scientific texts from the INSPEC database [7]. The guiding auxiliary data is composed of keywords chosen by the document authors. From these, the method is able to learn what is relevant in the full texts, which we measured with an expert categorization of the documents by informaticians.

Discriminative clustering improves the results of text clustering: The clusters are more closely related to relevant categories given by human experts, even though those categories were not used to train the models.

12.4 Discriminative clustering is vector quantization in learning metrics

We have shown [3] that at the limit of a large number of hard clusters (Voronoi regions), discriminative clustering performs a kind of vector quantization in learning metrics: The Euclidean distortion of normal vector quantization becomes replaced with the distortion computed in the new Fisher metric (12.1) defined by the conditional distributions $p(c|\mathbf{x})$ of the relevance-indicating auxiliary data. The Voronoi regions are, however, still defined in the original metric of the primary data space. This is due to the current parameterization of the clusters as the (soft) Voronoi regions of the original metric. In other words, the discriminative clustering algorithm is closely connected to the learning metrics principle.

References

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

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

[3] Samuel Kaski and Janne Sinkkonen. Principle of learning metrics for exploratory data analysis. Submitted to a journal.

[4] 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.

[5] Kimmo Kiviluoto. Predicting bankruptcies with the self-organizing map. *Neurocomputing*, 21(1–3):191–201, 1998.

[6] Kimmo Kiviluoto and Pentti Bergius. Analyzing financial statements with the self-organizing map. In *Proceedings of WSOM'97, Workshop on Self-Organizing Maps, Espoo, Finland, June 4–6*, pages 362–367. Helsinki University of Technology, Neural Networks Research Centre, Espoo, Finland, 1997.

[7] Jaakko Peltonen, Janne Sinkkonen and Samuel Kaski. Discriminative clustering of text documents. Submitted to a conference.

[8] Janne Sinkkonen and Samuel Kaski. Clustering by similarity in an auxiliary space. In Kwong Sak Leung, Lai-Wan Chan, and Helen Meng, editors, *Proceedings of IDEAL 2000, Second International Conference on Intelligent Data Engineering and Automated Learning*, pages 3–8. Springer, Berlin, 2000.

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

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

Chapter 13

Natural language modeling

Krista Lagus and Mikko Kurimo

13.1 Semantic analysis of Finnish words and sentences

Krista Lagus

The study of lexical organization, i.e. the study of how word-related information may be efficiently represented is important for many natural language applications, in particular, for smoothing language models. Furthermore, analysis of words as they are used in large corpora may reveal insight on the semantic properties of words that is relevant to linguistic research as well.

As shown in [1,6], the self-organizing map (Chapter 9) can be applied for clustering English word forms based on the words that have appeared in their immediate contexts. I 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 Englis, carry relevant semantic information [5]. 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 morphosyntactic features in the verb contexts (see [2,3]). Verbs were selected for study for two reasons: there exists a semantic reference classification of Finnish verbs [5] for comparing the results, and the semantic representation of verbs is considered an interesting problem in linguistics [4,5].

The corpus consisted of 17 million words of Finnish magazines and newspaper articles. In the first experiment, 25 verbs and their inflected forms were examined. In the second experiment, the 600 most frequent Finnish verbs were organized on a map, shown in Figure 13.1.

The results in both experiments show that even the simple features used, namely a set of morphological properties of the nearby words, collected over a large number of instances, were found to be suitable for obtaining automatically a semantic clustering and organization of verbs. When compared to a reference classification of Finnish verbs [5], the obtained 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.

Future research includes the study of verbs based on various other features (e.g. nouns in the context of the verb), and the organization of words from other syntactic categories.

References

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

[2] 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.

[3] K. Lagus. Studying similarities in term usage with self-organizing maps. In *Proceedings of NordTerm'01,* June 2001.

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

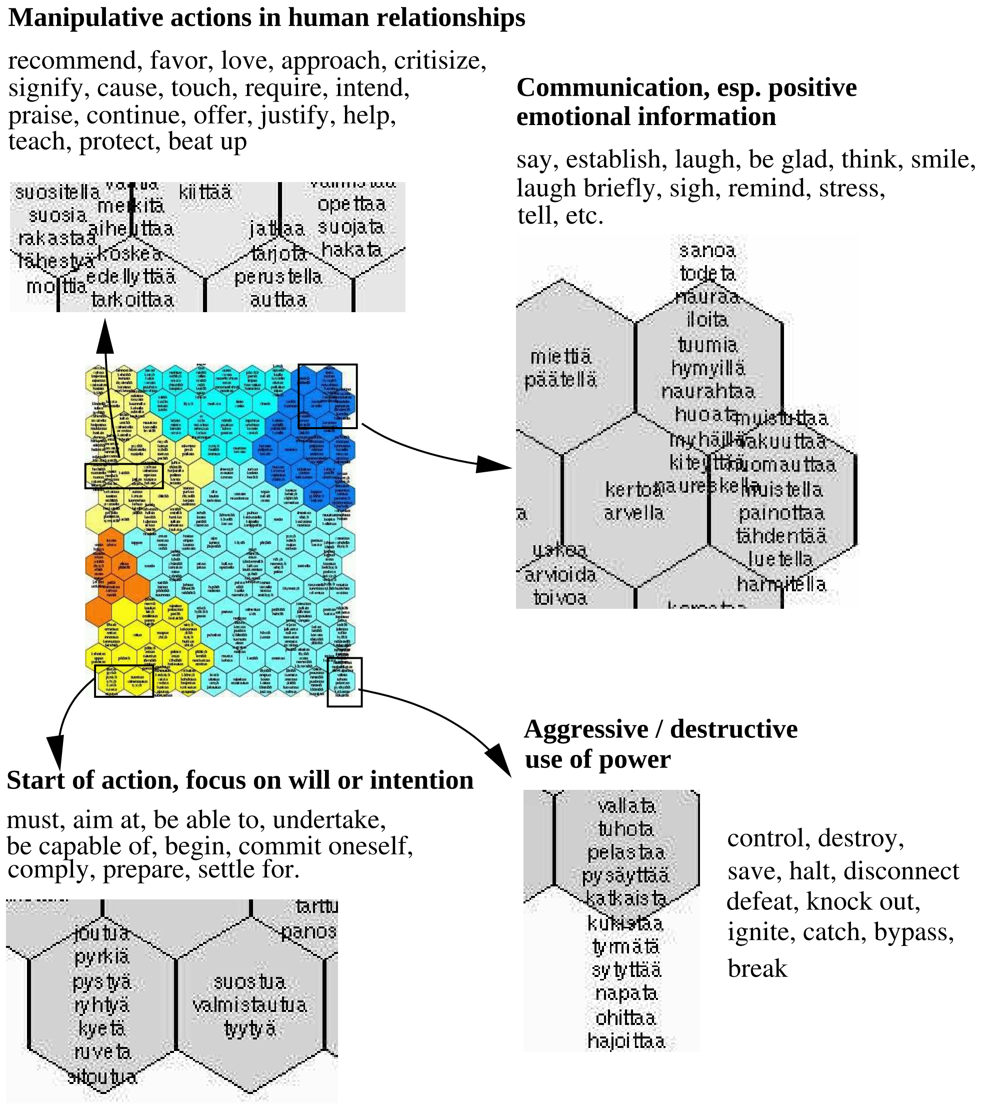


Figure 13.1: A map of the 600 most frequent verbs (base forms) in the Newspaper corpus. The verbs were organized on the basis of the distribution of morphological features in one preceding and two succeeding words, collected over all instances of the verb in any inflected form. The contents of four sample map regions are shown in the insets. In the reference classification (pp. 157–165 in [5], 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.

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

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

13.2 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 14.2) 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 [3] 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 [1] 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 10 or [2]). 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 13.2.

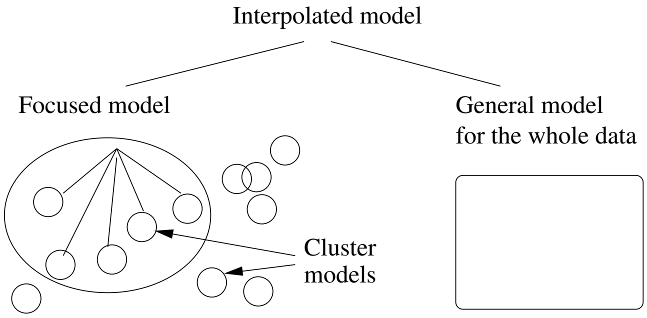


Figure 13.2: *A focusing language model obtained as an interpolation between topical cluster models and a general model.*

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.

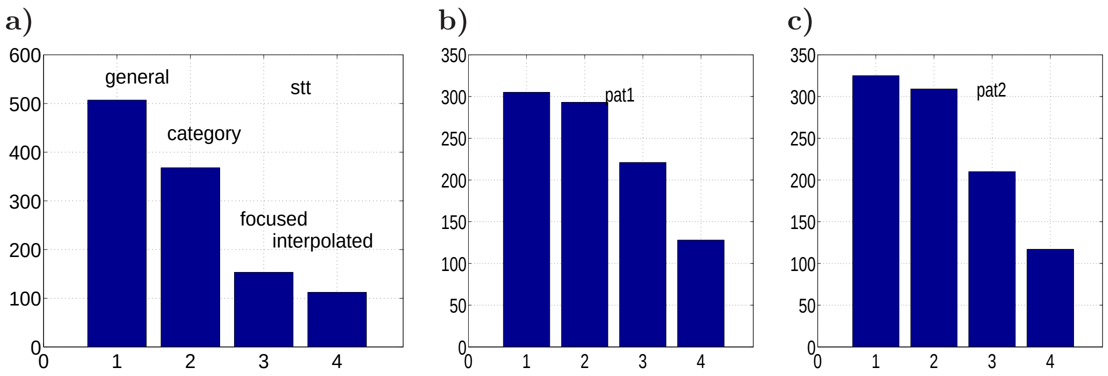


Figure 13.3: The perplexities of the different language models, **a)** for the Finnish STT news corpus, **b)** for smaller patent corpus and **c)** for larger patent corpus. The explanation of the bars in each figure, from left to right: 1. general model for the whole corpus, 2. category-specific model 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 13.3 indicate that the focusing model is superior in terms of perplexity when compared to a general "monolithic" trigram model of the whole data set. 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 next step in this project, for which the experiments are currently being performed, is to examine how well the obtained improvements in modeling translate to advancing speech recognition accuracy.

References

[1] 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.

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

[3] 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*, volume 1, pages 737–730, 2001.

Chapter 14

Speech recognition

Mikko Kurimo, Panu Somervuo, Vesa Siivola

14.1 Acoustic modeling

The general goal of automatic speech recognition (ASR) is to understand normal human speech and then to be able to perform some task based on this understanding. One application of ASR is a dictation system which converts spoken sentences into their written forms. In this sense the speech recognition can be defined as the mapping from the continuous acoustic signal to the discrete set of symbols.

There are several reasons for the difficulties in speech recognition. Natural speech has variations in many levels. In addition to that different speakers have different voices, there is also considerable variation in the voice of a single speaker. In normal conversation, some parts of the words may be emphasized more than others depending on the context. The loudness and the pitch of the voice may change and also the speaking rate may vary. Even if the speaker tries to use as steady a voice as possible, no two uttered sounds are generally equal. Therefore, in the speech recognition system, some limitations are usually made concerning the nature of the speech to be recognized. These limitations may include the number of the speakers, the size of the vocabulary, the amount of the noise in the speech, and the assumption that the input will always be speech.

Our projects in automatic speech recognition are aimed both to use the recognition system as a test bench for the neural network algorithms developed in the laboratory and to develop the system itself as a pilot application of the neural networks. Besides developing new **recognition algorithms**, we have also investigated new **acoustic features** and **context modeling**. Examples of the applications where we have used SOM and LVQ algorithms are shown in Table 14.1 and Figure 14.1.

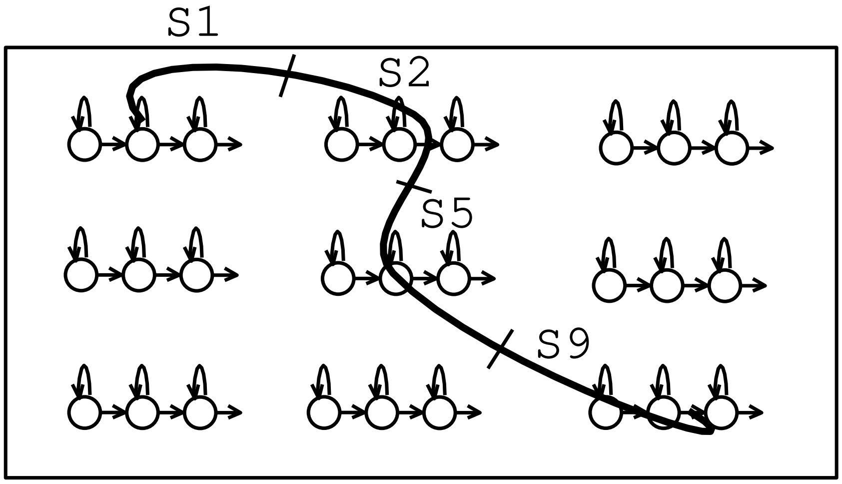


Figure 14.1: Competitive-learning segment models on the SOM. Each map node is associated with an HMM (with three states in this case) instead of a traditionally used single feature vector. The thick line represents the Viterbi segmentation of one input sequence. This corresponds to the best matching unit (BMU) search. The models of the BMUs and neighboring units are then updated by the corresponding segments.

The block diagram of a speech recognition system is shown in Figure 14.2. The recognition is based on connecting the hidden Markov models (HMMs) of the phonemes to decode the phoneme sequences of the spoken utterances. The output density function of each state in each model is a mixture of multivariate Gaussian densities. We have used the following scheme for the training of the models [1]. SOM is used first for initializing the phoneme-wise codebooks. Each model vector becomes then a mean vector of a

Gaussian kernel. After initialization, the training is continued by segmental-SOM or K-means algorithm. Segmental-LVQ is then applied for error-corrective training in order to obtain better phoneme discrimination. In the context of mixture density HMMs, we have developed methods for speeding up the recognition [1,3] based on the SOM structure.

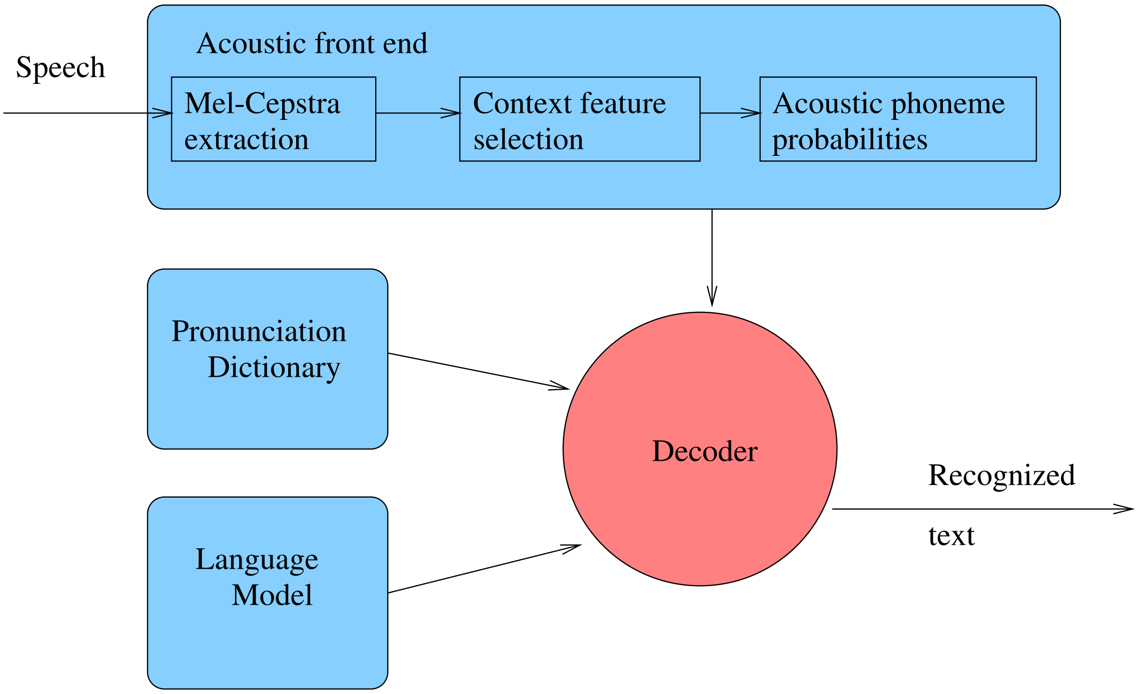


Figure 14.2: Overview of an ASR system.

Model associated with the SOM node:		Application:
1.	feature vector	kernel means of mixture densities [1]
2.	feature vector sequence	variable-length word templates [3]
3.	hidden Markov model	set of (non-linguistic) speech segments [3]
4.	symbol string	learning pronunciation dictionary [3]
5.	word n-gram	word cluster in a language model [2]
6.	word histogram	language model of a topic cluster [2]

Table 14.1: ASR-related SOM and LVQ applications.

As a promising future alternative for acoustic models in speech recognition an active research topic in the laboratory has also been the development of continuous state-space models of speech and Bayesian ensemble learning for latent variables. For more information on these topics see the chapter corresponding to the Bayesian modeling group.

Besides the recognition of speech, we have used our models also for the segmentation of new large Finnish speech corpora. The segmentation of the new data is an essential first step before the new material can be used for training. This work is related to the national USIX research program and has been helpful for the other participating Finnish speech research groups who are working on the same database.

14.2 Language modeling

The output of the phonemic vocabulary-free recognizer will inevitably contain some errors. Our current research is focused towards **large vocabulary continuous speech recognition** (LVCSR) systems and **language modeling**. The role of the language model is to control the search of the best phoneme or word sequence and improve the recognition. Since the best modeling methods are language specific, we cannot simply use the same models which have given good results e.g. for English. We have to cope with the special characteristics of the Finnish language which include e.g. a relatively free word order and a very large recognition vocabulary due to the number of inflected word forms and compound words. Some these problems are common to other non-english languages as well.

In order to better test the new methods and algorithms we are currently developing a new efficient decoder for the LVCSR task. This will be integrated in our speech recognition system.

One research topic has been how to better estimate the parameters of a language model. Since the models can consist of tens of millions of parameters, the parameter estimation is very sensitive to training methods and peculiarities of the training data. By carefully compressing the language model down to much fewer parameters, we increase the model's robustness and its ability to generalize for unseen case. One way to reduce the parameter count is to cluster similar words to one cluster and operate on these clusters instead of individual words [2].

An emerging new research topic is the use of the efficient language processing tools developed in the laboratory (WEBSOM) to organize language models based on the topical structure of the discourse [2]. 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 13.2).

References

[1] M. Kurimo. Using Self-Organizing Maps and Learning Vector Quantization for Mixture Density Hidden Markov Models. *PhD Thesis*. Helsinki University of Technology, Neural Networks Research Centre, 1997.

[2] 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*, volume 1, pages 737–730, 2001.

[3] P. Somervuo. Self-Organizing Maps for Signal and Symbol Sequences. *PhD Thesis*. Helsinki University of Technology, Neural Networks Research Centre, 2000.

Chapter 15

Methods for efficient utilization of SOM in data mining

Olli Simula, Jaakko Hollmén, Esa Alhoniemi, Juha Vesanto, Johan Himberg, Markus Siponen, Jussi Ahola, Juha Parhankangas, Mika Sulkava

The Self-Organizing Map (SOM) algorithm has been widely implemented in various software tools and libraries, for example, the SOM Toolbox [5]. However, for a common practitioner a difficulty with applying the SOM in data mining has been that there is no wide consensus or understanding of the methods needed for post-processing the SOM. Thus, methods for the two most important application areas of the SOM — visualization and cluster analysis — have been investigated and developed.

The goal of this work has been to enhance the data exploration process based on the SOM, see Figure 15.1. The 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. 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.

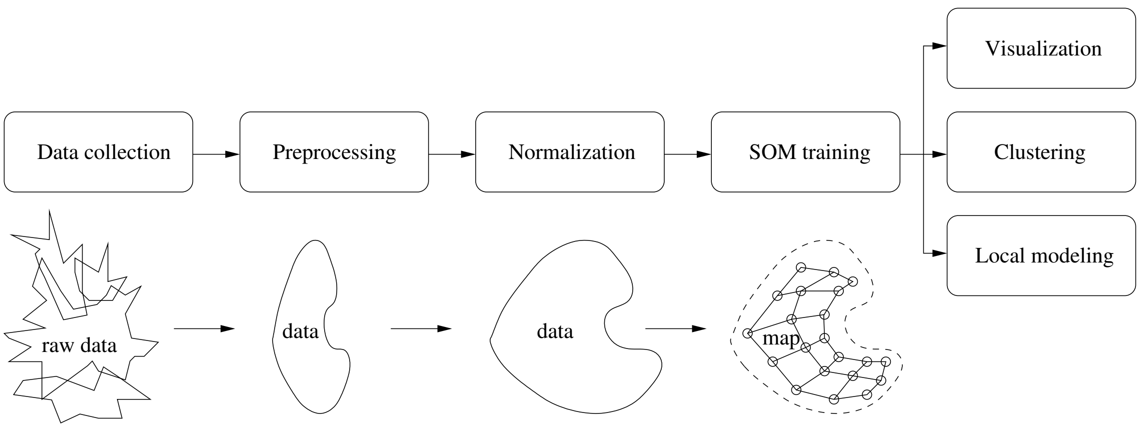


Figure 15.1: Applying the SOM in data mining. After data collection, the data is preprocessed, normalized, and a SOM is trained. In this work, we have concentrated on visualization and clustering in the post-processing of the SOM.

In [4], clustering of data using SOM is investigated. This is essentially a two-phase approach which reduces the computational load of clustering considerably, making clustering of large data sets feasible. Also, a SOM-based cluster visualization and its application to false coloring is described in a closely related paper [1].

Methods for interpretation of SOM clusters and a framework for (semi)automated analysis of hierarchical data are presented in [3]. Clusters are derived automatically, and then characterized by ranking the variables, and by constructing characterizing rules for the variables. The rules are formed by maximizing a novel measure of significance. In case of hierarchical data, the clusters form new variables for the upper level data, and the characterizations allow one to give them meaningful names.

An outline of a system for automatically generating data survey reports of table-format numerical data is described in [6]. Algorithms for constructing a cluster hierarchy and describing the clusters are proposed. Different methods and representations are combined to produce a unified and comprehensive report of the properties of the data manifold.

References

[1] J. Himberg. A SOM Based Cluster Visualization and its Application for False Coloring. In *Proceedings of International Joint Conference on Neural Networks (IJCNN2000)*, volume 3, pages 587–592, 2000.

[2] J. Himberg, J. Ahola, E. Alhoniemi, J. Vesanto, and O. Simula. Pattern recognition in soft computing paradigm. Volume 2 of FLSI soft computing series, ed. Nikhil R. Pal, chapter The Self-Organizing Map as a Tool in Knowledge Engineering. World Scientific, 2001.

[3] M. Siponen, J. Vesanto, O. Simula, and P. Vasara. An approach to automated interpretation of SOM. In N. Allinson, H. Yin, L. Allinson, J. Slack, editors, *Proceedings of Workshop on Self-Organizing Map 2001*, pages 89–94. Springer, June 2001.

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

[5] J. Vesanto, J. Himberg, E. Alhoniemi, and J. Parhankangas. SOM Toolbox for Matlab 5. Report A57, Helsinki University of Technology, Neural Networks Research Centre, Espoo, Finland, April 2000.

[6] J. Vesanto and J. Hollmén. An Automated Report Generation Tool for the Data Understanding Phase. In A. Abraham, M. Koeppen, editors, *Hybrid Intelligent Systems*, Advances in Soft Computing, Heidelberg, 2002. Physica Verlag. In print.

Chapter 16

Data-based analysis of industrial processes

Olli Simula, Miki Sirola, Jaakko Hollmén, Ignacío Diaz, Sampsa Laine, Esa Alhoniemi, Johan Himberg, Jussi Ahola, Jukka Parviainen, Timo Similä, Jarkko Tikka

16.1 Extraction and identification of operational states of a paper machine

An underlying assumption in most data analysis, monitoring and diagnostics methods is that the process considered is in a stationary state. However, in practice this assumption is often ignored due to lack of tools for the analysis of operational states. Ignoring the stationarity assumption results in poorer performance of the monitoring and diagnostic methods. In paper production, there exist many operational states that affect the performance of the process, yet are not clearly distinguishable from measured data.

When a monitoring or diagnostics system for improved fault detection is implemented, a problem frequently arises: are the data used as reference for future process operation collected from a stationary state and is this state achievable repeatedly? Also, the stability of the process – and thus the effect of a same kind of disturbance – may greatly vary depending on various circumstances, which are not connected to the origins of the disturbance. Thus, another question concerning the states of a process is: if the same product can be produced in various ways, which is the most profitable?

In order to study and manage different process states, a cluster analysis tool was developed. It consists of all the modules that are required in the cluster analysis: preprocessing, clustering and interpretation parts. The whole framework with a description of the tool is reported in [1]. A method for interpretation and comparison of clusters – that are in this context seen as operational states of a process – was also reported separately in [2]. Some visualizations of one process state are shown in Figure 16.1.

References

[1] M. Federley, E. Alhoniemi, M. Laitila, M. Suojärvi, and R. Ritala. State management for process monitoring, diagnostics and optimization. *Pulp & Paper Canada* (to appear).

[2] E. Alhoniemi and O. Simula. Interpretation and Comparison of Multidimensional Data Partitions. In Michel Verleysen, editor, Proceedings of the 9th European Symposium on Artificial Neural Networks, pages 277–282. D-Facto, 2001.

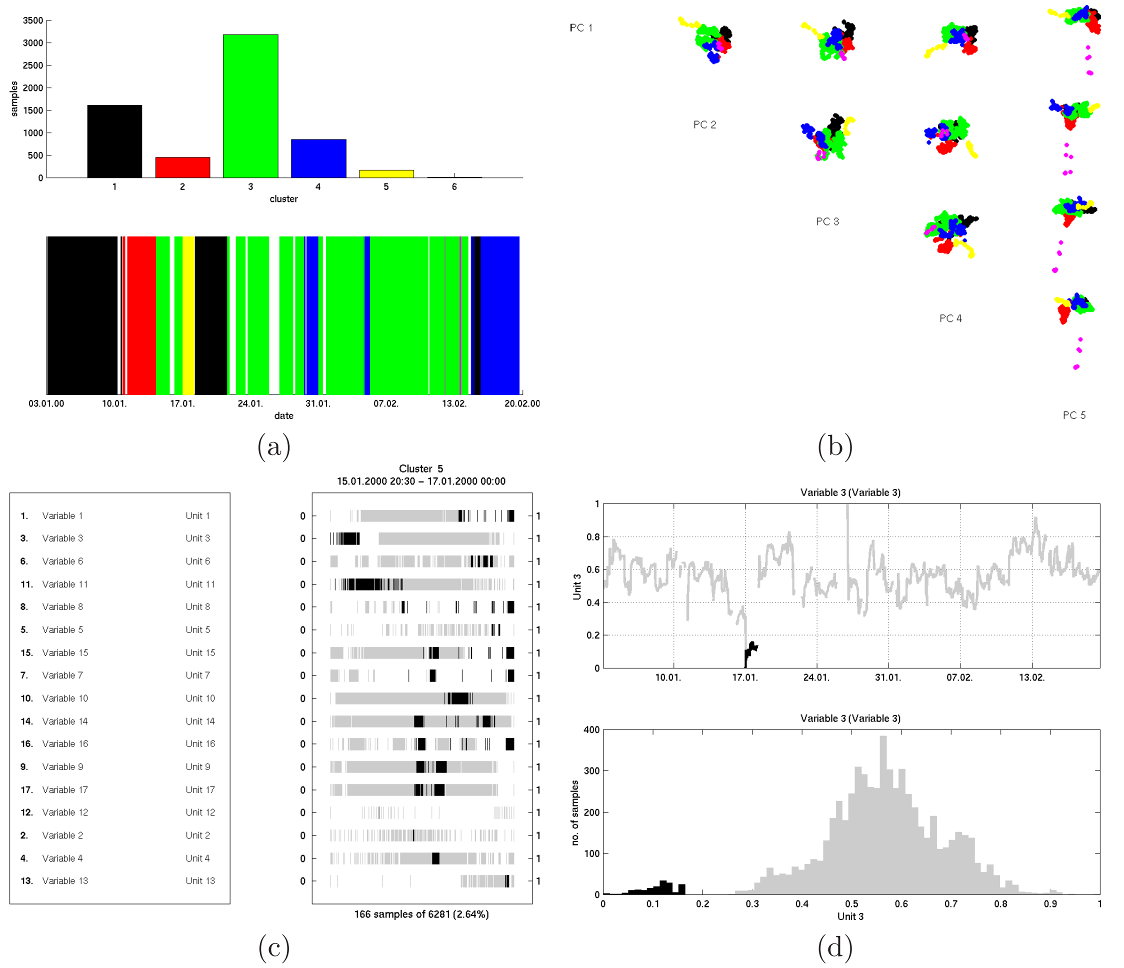


Figure 16.1: Different visualizations produced by the clustering tool. (a) Sizes of six extracted clusters and their location in time, (b) PCA projection of the clusters, (c) a "thumbnail" display of cluster 5 (which are marked by yellow color in (a) and (b)), and (d) a time series and histogram display of one variable (number 3) that is important in characterization of cluster 5.

16.2 Variable selection, feature extraction and visualization in data based process study

Consider yourself as a process engineer with a task of solving a recurring process control problem, for example, the disturbances frequently caused by the changes in the process feed type. The natural starting point is to collect information by interviewing process operators, discussing with colleagues and gathering process data. Then you would return to study the acquired information to define the problem, to define a solution, and to create an application to support process control. This project studies how the tools of variable selection, feature extraction and data visualization can support this task.

In this project, a full-scale mineralogical and a pilot scale pharmaceutical process were considered. In [1] manual variable selection and feature extraction were considered. The self-organizing map (SOM) was used to visualize the process state in [2]. An algorithmic variable selection approach is reported in [5]. Based on this, it is demonstrated how to find the on-line variables that best describe an off-line defined process phenomenon [3]. Another use for the variable selection technique is reported in [4], where the process control variables best explaining changes in process control success were searched for. The current work aims at creating a methodology for finding the variables that train a SOM in which the phenomenon defined by the user is presented as clear as possible.

References

[1] S. Laine, K. Pulkkinen, and S.-L. Jämsä-Jounela. On-line determination of the concentrator feed type at Outokumpu Hitura Mine. *Minerals engineering* pages 881–895, No. 8–9, August 2000.

[2] J. Rantanen, S. Laine, O. Antikainen, J.-P. Mannermaa, O. Simula, and J. Yliruusi. Visualization of pharmaceutical unit operations using self-organizing maps (SOM). J. Pharm. Biomed. Anal., pages 343–352, January, 2001.

[3] S. Laine. Combining Off-Line and On-Line Information in Process Study Using the Self- Organizing Map (SOM) Proceedings of IEEE Mountain Workshop of Soft Computing in Industrial Applications, ed. Embrechts J., VanLandingham H., Ovaska S., Blacksburg, USA, pages 71–76, June 2001.

[4] S. Laine. State Based Process Study Using the SOM and a Variable Selection Technique in Evolving Solution with Neural Networks. In Proceedings of International Conference on Engineering Applications of Neural Networks, ed. R. Baratti, J. Caete. Cagliari, Italy, pages 15–22, July 2001.

[5] S. Laine. Finding the Variables of Interest. *Minerals engineering* (accepted).

16.3 From fault detection to fault identification with normal condition process data

Fault detection and fault identification are the two problem types encountered when analyzing industrial process data. Solving the fault detection problem in the presence of normal condition data only — in the framework of novelty detection — one can train a process model of the normal state and detect any deviations from this model as being abnormal. On the contrary, if fault data is available, fault identification problem can be solved in the classification setting, where process data are the symptoms and the faults are the classes. One interesting problem class that often occurs in practice is the need to solve the fault identification problem with normal condition data only, and perhaps a few labeled samples of the faults or some general prior knowledge of the fault types. We are interested in solving this problem that in fact falls in between the standard treatments of fault detection and fault identification problems.

In [1], we compare trained models from normal process condition data by studying the usefulness of residuals in the problem of fault identification. More specifically, we study the generation and visualization of residuals for detecting and identifying novel, unseen faults using autoassociative models learned from process data. Least squares and kernel regression models are compared on the basis of their ability to describe the support of the data. Theoretical derivations are used to show that kernel regression models are more appropriate in approximating the normal condition model. Moreover, experiments on vibration and current data from an asynchronous motor confirm the theory and yield more meaningful results when testing the models with some known fault classes.

References

[1] I. Díaz and J. Hollmén. Residuals Generation and Visualization for Understanding Novel Process Conditions. Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN'02) (to appear).

16.4 Decision models for computerized decision support

Several 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 were the nuclear power plants. Decision support in both the control room and maintenance were covered. The methodological frame of the models consists of decision theory, knowledge engineering and object-oriented modeling. One of the models, the use of decision analysis methodology in maintenance problems, is presented in [2].

The problem in the control room is how to produce instructions for the personnel responsible for the plant in situations where faults have occurred at various stages. In maintenance mainly off-line support is needed. Methodology and models based on application knowledge are developed to meet the functional requirements of decision support systems. Prototypes are built to verify the usability of the methodologies and models. The tools are demonstrated and tested by solving concrete decision problems. One of the key issues is the methodology developed in strategy planning, including both generation and selection of strategies.

After these works mentioned 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. As results of these studies tested models and concepts are got, but the evaluation of them is difficult, because no competent measures exist 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. The methodologies used in Finnish industry, including chemical industry, nuclear and conventional power production, has been collected and analyzed in an interview study. A decision model for state monitoring has been developed. This model utilizes fuzzy methodologies, traditional calculation techniques and heuristics.

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]. Comparisons 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.

References

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

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

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

16.5 The other projects

Fault detection in continuous casting. Continuous casting of steel has been investigated in a project with Suomen Perusmetalli. The casting process is a large scale industrial method for producing steel slabs for further refinement, e.g., for hot rolling. During the casting, surface ruptures may appear on the steel slab. They degrade the product quality, and sometimes they may even cause a breakthrough of liquid steel that leads to a long maintenance break at the casting plant. The measurements monitoring the casting process are done using thermocouples inside the mold. These temperature measurements have been analyzed using SOM-based visualization tools in order to find efficient features for rupture and breakthrough warnings [1]. At the moment, satisfactory features have been discovered and used to build a classifier to detect ruptures. The state of the process can be indicated using the trajectory of measurements on the visual SOM display. The SOM can, thus, be used for on-line monitoring of the casting process.

Analysis of hot rolled steel production. During the period 1.2.1997–31.1.2000 "Application of Neural Network based Models for Optimisation of the Rolling Process", NEUROLL, was carried out. The project was financed by the European Union under the program Brite-Euram III. The rationale of this project was to apply intelligent computation methods, especially neural networks to the steel strip rolling, the ultimate goal being to improve quality and throughput of the products. From the product point of view, the project was concentrating on such dimensional properties as width, thickness and flatness; on the material properties, especially phase transition; and on the surface quality control.

The research carried out in the NEUROLL project was focused on two subprojects: (1) The detection of complex relationships between measured or calculated process parameters and input/output variables and (2) analysis of the surface quality of the hot rolled steel strips. The SOM based data analysis was carried out for visualization and classification of various measurements and data dependencies. The study has concentrated on finding correlations between quality parameters (e.g., width and thickness deviation and surface defects) and other process parameters using traditional correlation analysis and methods based on the SOM. [3, 2] As an example, Figure 16.2 below shows the dependence of the quality parameters on the chemical content, including the temperature measured after the last stand, could be seen that the deviation of the temperatures has a clear influence on the thickness and width deviations. [4]

References

[1] J. Himberg, J. Ahola, E. Alhoniemi, J. Vesanto, and O. Simula. Pattern recognition in soft computing paradigm. Volume 2 of FLSI soft computing series, ed. Nikhil R. Pal, chapter The Self-Organizing Map as a Tool in Knowledge Engineering. World Scientific, 2001.

[2] J. Parviainen. Data Mining for Finding Surface Defects in Steel Strips. Master's thesis, Helsinki University of Technology, 2000.

[3] J. Ahola. Analysis of Mechanical Properties of Hot Rolled Steel Strips Using SOM-based Data Mining. Master's thesis, Helsinki University of Technology, 2001.

[4] Application of Neural Networks based Models for Optimisation of the Rolling Process (NEUROLL). Final Technical Report, Contract No. BRPR-CT96-0342, Project No. BE96-3061.

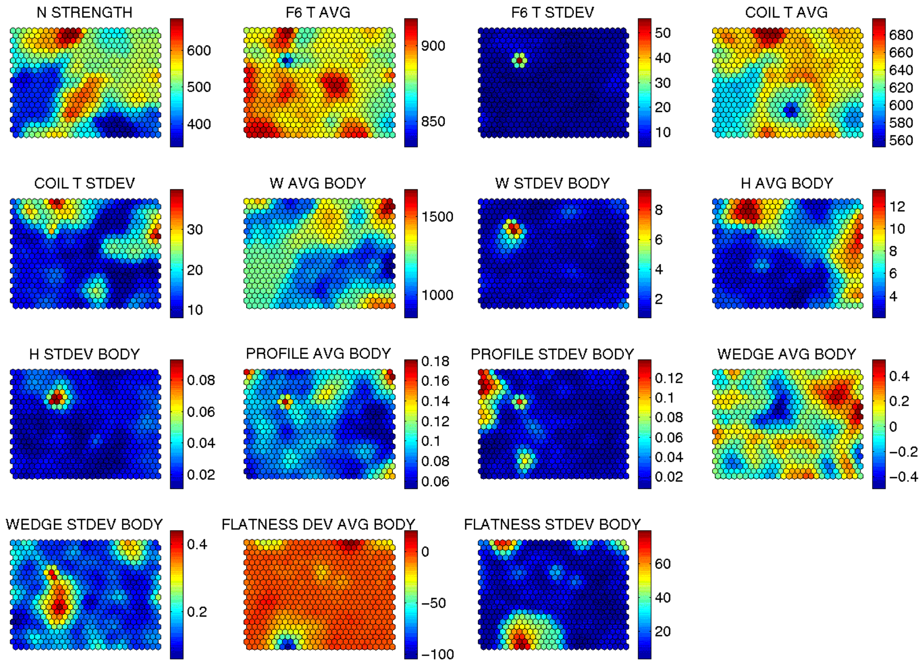


Figure 16.2: Temperature vs. quality parameters (micro-alloyed steels).

Chapter 17

Data-driven analysis of telecommunication systems

Olli Simula, Kimmo Raivio, Jaakko Hollmén, Kimmo Hätönen, Sampsa Laine, Pasi Lehtimäki, Timo Similä

17.1 Introduction

Large amounts of data are generated by the telecommunications systems in order to log events for later analysis in problem cases or to charge for the services used. We have applied our earlier experience in studying the industrial processes in the telecommunications context: the data-driven approach may be applied equally well in this context as in the analysis of industrial processes.

17.2 Analysis of mobile radio access network

The 3G cellular systems will offer services beyond the capabilities of today's networks. The variety of services requires some modifications on the radio network planning and optimization process. One of the modifications is related to the quality of service (QoS) requirements and control. For each service the QoS targets have to be set and naturally also met. With configuration parameters optimum operation point for each cell should be found.

In this project, wideband code division multiple access (WCDMA) mobile network has been analyzed using the Self-Organizing Map [3] [2]. The goal is to develop efficient adaptive methods for monitoring the network behavior and performance. Special interest is on finding clusters of mobile cells, which can be configured using similar parameters. The data has been generated using a WCDMA radio network simulator [1].

Figure 17.1: Model used in classification of mobile cells.

The method utilizes the SOM algorithm both in feature extraction and clustering the cell features. A block diagram of the method is shown in Figure 17.1.

The data vectors of all the cells are clustered using a two phase clustering algorithm [4]. In that algorithm, a Self-Organizing Map is trained using 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 clusters can be defined. The input data vectors are classified using labeled SOM codebook vectors.

Next, some results of the network analysis are presented. Five variables are used in the analysis and two of them describe the QoS level in the network. Clusters of SOM codebook and some automatically generated rules of the clusters are shown in Figure 17.2a and 17.2b.

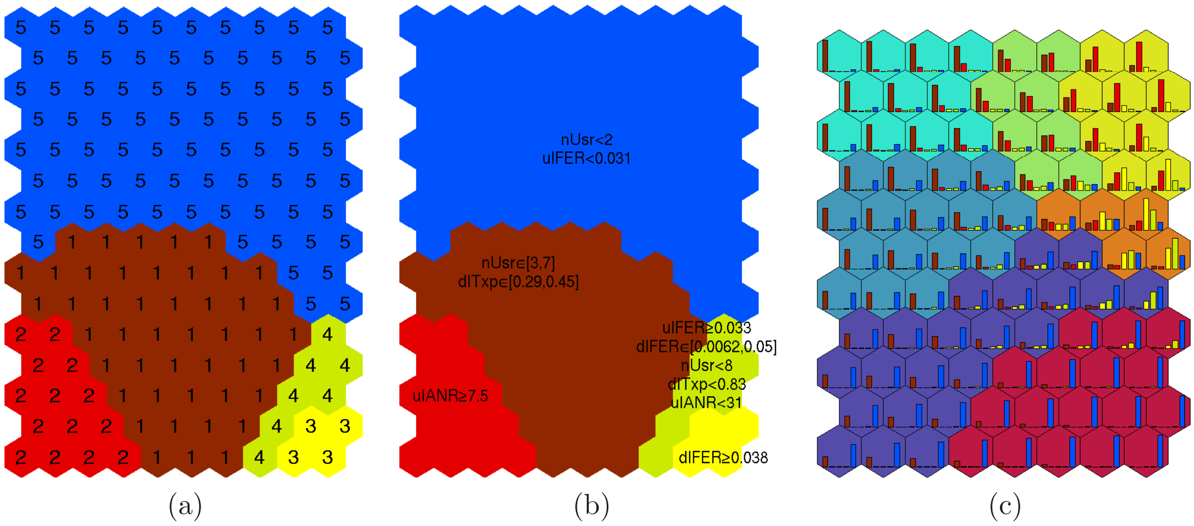
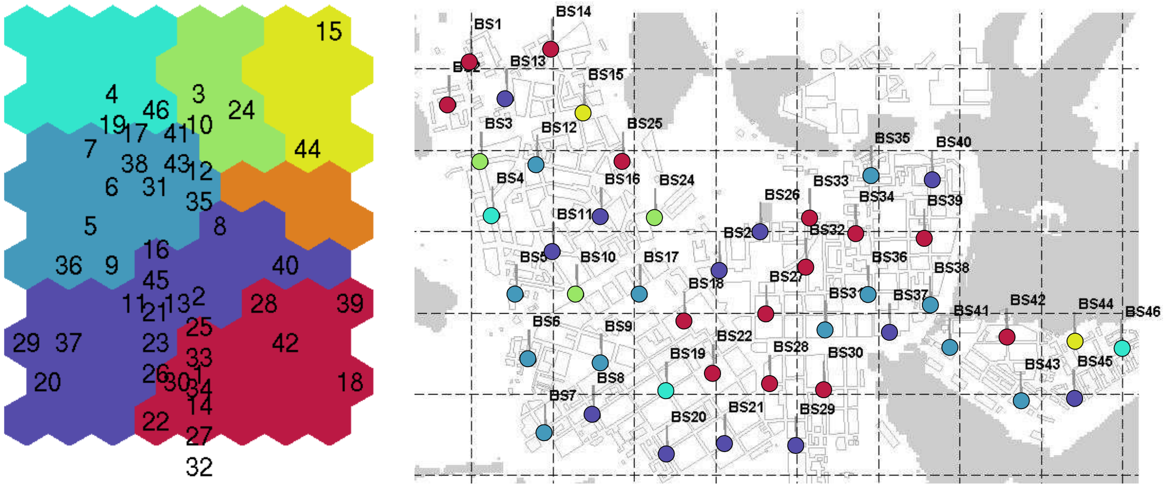


Figure 17.2: Clusters of the map of data vectors (a), rules of clusters (b) and clusters of histogram map (c).



(a) (b)

Figure 17.3: BMUs of mobile cells on histogram map (a) and locations of classified cells (b).

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 clusters of histogram SOM are shown in Figure 17.2c and the best-matching units (BMU) of the mobile cells in Figure 17.3a. The classified cells and their locations are presented in Figure 17.3b.

In this method, two level clustering procedure has been used because of high data rate. In 3G systems, the sampling rate can be from ten to hundred samples per second. Similar methods can be used also in lower sampling rate systems like GSM, but then it is enough to process the data by using only one clustering level.

References

[1] S. Hämäläinen, H. Holma, and K. Sipilä. Advanced WCDMA radio network simulator. In *Personal, Indoor and Mobile Radio Communications*, volume 2, pages 951–955, Osaka, Japan, September 12-15 1999.

[2] P. Lehtimäki, K. Raivio, and O. Simula. Mobile radio access network monitoring using the self-organizing map. In *Proceedings of European Symposium on Artificial Neural Networks*, Bruges, Belgium, April 24 - 26 2002. (accepted).

[3] K. Raivio, O. Simula, and J. Laiho. Neural analysis of mobile radio access network. In *IEEE International Conference on Data Mining*, pages 457–464, San Jose, California, USA, November 29 - December 2 2001.

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

17.3 The use of preprocessing and visualisation techniques to analyze telecommunications data

The GSM and third generation cellular phone networks produce masses of data. As the networks develop fast, the operating personnel can not keep up-to-date with the new conditions of the system. They have problems to evaluate the data: to select the most important variables and cells for study, and to assess the state of the network. This study searches for the preprocessing and visualization techniques that support inexperienced process operators. The display designed to given an overview to the GSM-network is presented in Figure 17.4. The nodes of the graph are the GSM-transmitters; the arcs indicate cells. The color of the node indicates the status of each GSM-cell.

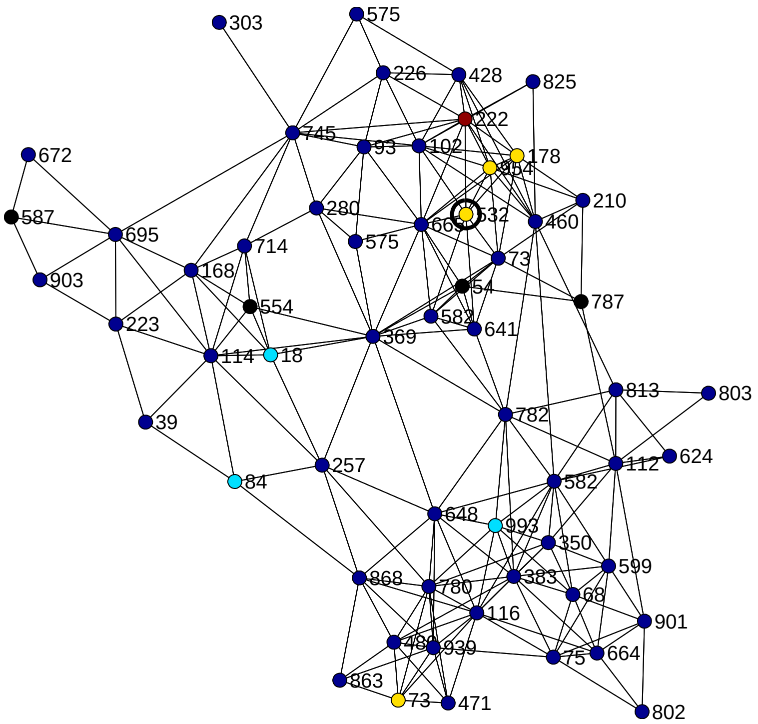


Figure 17.4: A graph depicting the status of a GSM-network.

The colors are a result of a data analysis process. The process is based on the proper value range of each variable defined by an expert into fuzzy membership functions. These membership functions are used to pretreat the data: to scale the variables and to remove outliers. The resulting data is used to train a SOM, which is clustered [1] to find the typical states of the GSM-cells. The clusters are allocated a color known to the user. For example, red color can be used to indicate problems.

The approach has been created together with a senior expert of the GSM-networks in the Nokia company, Mikko Toivonen. The research is based on the funding of the Suomen Akatemia and the Nokia Foundation.

References

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

17.4 Fraud detection in mobile communications networks

Telecommunications fraud, the illegitimate use of services, presents a financial burden to the network operators, since the service charges from the services used by fraudsters can not be collected. In fraud detection, the goal is to discover the presence of illegitimate calling activity of telecommunications customers. What makes calling activity illegitimate in the above sense is that the subscriber's intention at the time of calling is not to pay for the services used. The intentions can not be observed directly, but they are reflected in the calling behavior. The calling behavior is collectively described by the call data, which in turn can be observed. Therefore, it is reasonable to use call data as a basis for subsequent analysis, both in formulation of models through learning and in detection of fraudulent behavior.

Whereas the call data exemplifies individual patterns of normal or fraudulent behavior, the goal is to formulate a *model* so that it would essentially articulate the same thing as data but on a more abstract and general level. The approach taken in this work is to *learn* a representation of normal and fraudulent behavior that can be used in decision making. For this end, probabilistic models and neural networks have been applied [1]. These models handle uncertainty in the domain in a favorable fashion and they can process the abundance of data present in the telecommunication domain. In this work, we have applied dynamical modeling of behavior in fraud detection. This has proved to be a promising line in fraud detection, and in user profiling and classification in general, since behavior at the most natural level is defined in a temporal context.

User profiling and classification for fraud detection is reported in [1] (see also the section on doctorate theses).

References

[1] J. Hollmén. *User Profiling and Classification for Fraud Detection in Mobile Communications Networks*. PhD thesis, Helsinki University of Technology, 2000.

Chapter 18

Other projects

18.1 Mixture density from autonomous experts

Jarkko Salojärvi and Samuel Kaski

In mixture density modeling it is assumed that each data sample is generated by one of a set of independently operating mixture components, the whole data set being a mixture of samples from them. When fitting a mixture of models or "experts" to data, the modeling task is broken into partially overlapping subtasks assigned to relatively simple experts. This increases analytical tractability and makes the results more easily understandable, a topic especially important in tasks where the goal is to summarize large data sets and gain understanding of their essential characteristics.

Although each expert generates data independently, an expert within a mixture cannot be fitted autonomously to data because the contributions of all the other experts must be known. In [1,2] we introduce a variant of the mixture model in which the experts can learn autonomously by maintaining simplified, periodically updated representations of the other experts. If the representations are simpler than the original experts themselves, the complexity of each expert will be considerably lower than that of the whole mixture. By assigning the autonomously operating experts to different processors, computational load can be distributed. The autonomous learning procedure can be interpreted so that each expert modulates its learning by its estimate of the probability that it is responsible for the data.

The principle is demonstrated with toy data in Figure 18.1. We have applied the autonomous learning to several data sets [2].

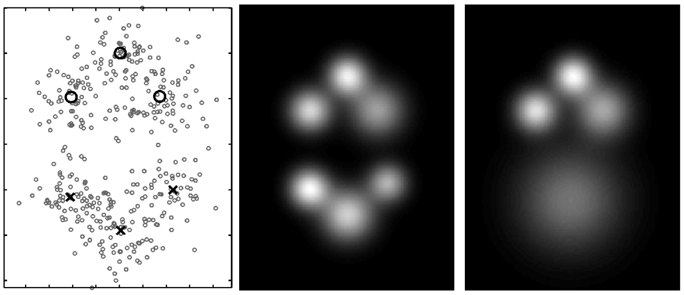


Figure 18.1: A toy example demonstrating the autonomous experts. Left: The data (small circles) was modeled with two autonomous experts, each consisting of three Gaussians. The means of the Gaussians are shown with circles for expert 1 and crosses for expert 2. Center: the density generated by the mixture. Right: the density plot shows how the first expert sees the data. The large Gaussian at the bottom represents the other expert.

References

[1] S. Kaski and J. Salojärvi. Generative mixture modeling by autonomous estimators. In N. Baba, L. C. Jain, and R. J. Howlett, editors, *Knowledge-Based Intelligent Information Engineering Systems & Allied Technologies, Proceedings of KES'2001*, volume 1, pages 250–254. IOS Press, Amsterdam, 2001.

[2] J. Salojärvi and S. Kaski. Mixture density from autonomous experts. *International Journal of Knowledge Based Intelligent Engineering Systems*, 6(1):48–55, January 2002.

18.2 Self-organizing map-based information visualization is trustworthy

Jarkko Venna and Samuel Kaski

One of the main uses of Self-Organizing Maps is information visualization in exploratory data analysis. A Self-Organizing Map of a large unknown multivariate data set gives a first comprehensive overview of the similarity relationships of the data samples and cluster structures in the set.

Several measures have been proposed for comparing nonlinear projection methods but so far no comparisons have taken into account one of their most important properties, the trustworthiness of the resulting neighborhood or proximity relationships. In visualizations it is crucial that the visualized proximities can be trusted upon: If two data samples are close to each other on the display they should be close-by in the original space as well. We have proposed to measure trustworthiness by the number of samples arriving from outside of the original neighborhood to a small neighborhood on the display (more accurately: the rank distance from the neighborhood [1]).

A sample experiment is shown in Figure 18.2**a**. For the important small neighborhoods (small k), neighborhood relationships visualized by the Self-Organizing Map (SOM) and its variant, the Generative Topographic Mapping (GTM), are more trustworthy than visualizations produced by traditional multidimensional scaling-based nonlinear projection methods. Yet, Self-Organizing Maps preserve the original neighborhoods with a comparative accuracy (Fig. 18.2**b**).

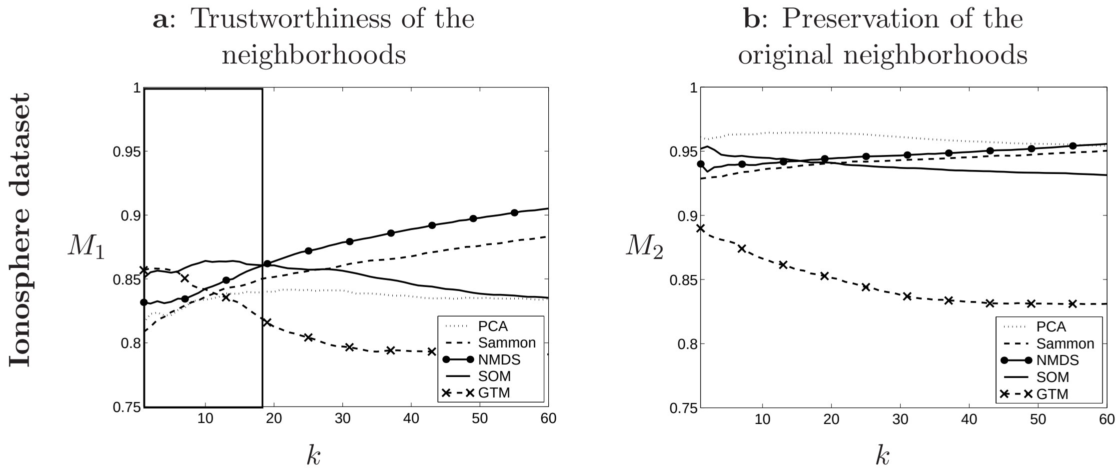


Figure 18.2: Trustworthiness (M_1) of the neighborhoods after projection (**a**) and preservation (M_2) of the original neighborhoods (**b**) for a sample data set as a function of the neighborhood size k. The SOM outperforms the MDS-based methods in the boxed region. PCA: principal component analysis, Sammon: Sammon's mapping, NMDS: non-metric multidimensional scaling.

References

[1] J. Venna and S. Kaski. Neighborhood preservation in nonlinear projection methods: An experimental study. In G. Dorffner, H. Bischof, and K. Hornik, editors, *Artificial Neural Networks—ICANN 2001*, pages 485–491. Springer, Berlin, 2001.

18.3 SOM in detecting states of sleep and wakefulness in polysomnographic data of mentally retarded persons

Sirkka-Liisa Joutsiniemi, Maija-Liisa Laakso, Lea Leinonen, Jussi Nurminen, Teuvo Kohonen

Interview and observation based studies show that sleep problems are common in the mentally retarded [1], [3]. Polysomnography is 'the golden standard' for all objective methods in sleep research. There are often difficulties, however, in applying this method in the investigations on subjects with damaged brains. The features used in the classification of wake and sleep stages of healthy subjects may be lacking or deformed, or there may be abnormalities disturbing the definitions. The patient group is heterogeneous and no common rules apply to them.

In collaboration with Rinnekoti Foundation, we are developing a SOM-based method to help the clinicians in the analysis of the polysomnography of the mentally retarded. The ability of the SOM to learn unsupervised the regularities in the individual data and visualize the structure of it offers a good basis for such a tool [2] (Fig. 18.3). A prerequisite for efficient SOM are new signal features suitable for detecting wake and sleep states in the mentally retarded. The work concerning the feature extraction is going on.

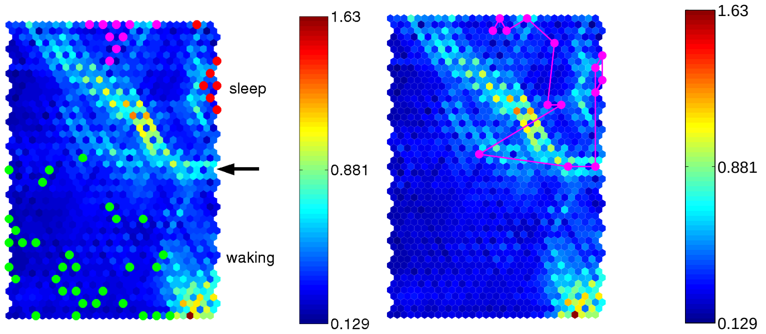


Figure 18.3: The SOM diagrams of the 24-h polygraphy of a healthy subject. *Left:* The border between waking and sleep areas is indicated by an arrow. On the sleep area, the red dots on the right are projections of 60 s deep sleep (S4) episodes of the polygraphy, dots on the left of the same area are projections of REM sleep episodes. The lower part of the map receives projections of waking (green dots). *Right:* The trajectory of the moving projection of transition from S4 to REM sleep.

References

[1] Lindblom N, Heiskala H, Kaski M, Leinonen L, Nevanlinna A, Iivanainen M, Laakso M-L: Neurological impairments and sleep-wake behaviour among the mentally retarded. *Journal of Sleep Research 10, 309-318, 2001.*

[2] S.-L. Joutsiniemi, J. Nurminen, T. Kohonen SOM in detecting states of sleep and wakefulness in polysomnographic data. *Proceedings of the Fourth International Conference on NNESMED 2001*, pages 97–101.

[3] 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 19, 1-19, 2002 (in press).*

18.4 Independent variable group analysis

Krista Lagus, Esa Alhoniemi, Harri Valpola

The goal of unsupervised learning is to extract an efficient representation of the statistical structure implicit in the observations. A good model is both accurate and simple in terms of model complexity, i.e., it forms a *compact representation* of the input data.

In problems with a large number of diverse observations there are often groups of input variables that have strong mutual dependences within the group but which can be considered practically independent of the input variables in other groups. It can be expected that the larger the problem domain, the more independent groups there are. Estimating a model for each independent group separately produces a more compact representation than applying the model to the whole set of variables. Compact representations are computationally beneficial and, moreover, offer better generalization.

We present an approach called independent variable group analysis (IVGA), where the dependences of variables within a group are modeled, whereas the dependences between the groups are neglected [2]. Usually such variable grouping is performed by a domain expert, prior to modeling with automatic, adaptive methods. As expert knowledge may be unavailable, or expensive and time-consuming, automating the task can considerably save resources. The IVGA is a practical, efficient and general approach for obtaining compact representations that can be regarded as sparse codes, as well.

The $\mathrm{IVGA}_{\mathrm{VQ}}$ algorithm

Any IVGA algorithm consists of two parts, (1) grouping of variables, and (2) construction of a separate model for each variable group. An independent variable grouping is obtained by comparing models with different groupings using a suitable cost function. In principle any model can be used, if the necessary cost function is derived for the model family.

We have applied the IVGA approach for a situation where the dependences within variable groups are modeled using vector quantization (VQ), and derived the necessary cost function for model optimization with VQ. For optimization of the model we use the variational EM-algorithm (cf. e.g. [1,3]).

Since it is generally computationally prohibitive to try all the different groupings of D variables into any number of groups, some heuristic optimization strategy has to be utilized. We have used the following strategy:

- **Initialize:** Assign each variable to its own group. Model each group using VQ and calculate total model cost.
- **Repeat until time limit or model cost limit:**
  - Consider a change in the grouping of variables:
    - Move a variable from group to another
    - merge two groups
    - Run $\mathrm{IVGA}_{\mathrm{VQ}}$ recursively for a variable group or the union of two groups (split or merge+split)
  - Model each group using a VQ and calculate total model cost. If cost improved, accept the change.

As a cost function one can use negative log-likelihood of the data given the model, namely $-\ln p(\mathbf{x}|H)$. The total model cost L_{tot} needed for comparing variable groupings is the sum of costs of individual variable groups $L_{\text{tot}} = \sum_g L_g = \sum_g - \ln p(\mathbf{x}_g|H_g)$, where g is the index of a group of variables, and x_g and H_g are the data and the model related to that variable group, respectively. Regarding the VQ model and the variational EM algorithm used for optimization of the model parameters the interested reader is referred to [2].

Experiments

Experiments were carried out to (1) verify the general IVGA principle by comparing the obtained groupings to a known categorization of variables, and (2) to study the performance of the presented algorithm by comparing model costs obtained using IVGA_{VQ} and regular VQ.

The data set consisted of 1000 images from the PicSOM project (see Chapter 7). Three kinds of features (variables) had been calculated to represent each image: FFG, RGB, and texture. It is reasonable to assume that dependences between variables in different categories are weak (cf. e.g. [4]). The original images were represented by 144 features, of which 50 were randomly chosen here: 27 FFT features (feature set A), 7 RGB features (feature set B), and 16 texture features (feature set C).

A summary of the results is shown in Table 18.1. The VQ(A+B+C) denotes the running of the regular VQ for the combined set of features A, B, and C. In particular, the results show that IVGA_{VQ}(A+B+C) obtained a clearly better model than plain VQ(A+B+C).

Table 18.1: Summary of the results of the experiments. The groupings obtained using IVGA_{VQ} contain variables from one category only.

Experiment	**Total cost**	**# VQs**	**# Params**
1. VQ(A+B+C)	-138115.5	1	2200
2. VQ(A) + VQ(B) + VQ(C)	-145796.2	3	1045
3. IVGA_{VQ}(A+B+C)	-147206.8	12	618
4. IVGA_{VQ}(A) + IVGA_{VQ}(B) + IVGA_{VQ}(C)	-147934.3	9	712

In conclusion, the experimental results show that it is worthwhile to group variables according to independence, and that the presented algorithm is able to do this and in doing so, obtains more compact models.

References

[1] G. E. Hinton and D. van Camp. Keeping neural networks simple by minimizing the description length of the weights. In *Proceedings of the COLT'93*, pp. 5–13, Santa Cruz, California, USA, July 26–28, 1993.

[2] K. Lagus, E. Alhoniemi, and H. Valpola. Independent Variable Group Analysis. In *Proceedings of International Conference on Artificial Neural Networks - ICANN 2001*, pp. 203–210. Vienna, Austria. Springer, 2001.

[3] H. Lappalainen and J. W. Miskin. Ensemble learning. In M. Girolami, ed., *Advances in Independent Component Analysis*, pp. 76–92. Springer-Verlag, Berlin, 2000.

[4] E. Oja, J. Laaksonen, M. Koskela, and S. Brandt. Self-organizing maps for content-based image database retrieval. In E. Oja and S. Kaski, eds., *Kohonen Maps*, pp. 349–362. Elsevier, 1999.

18.5 Modeling visual attention

Teuvo Kohonen

Introduction

The term *attention* has been used in connection with many different psychological, behavioral, and physiological states and actions. Here we may simply define attention as a set of those cognitive functions by which primary sensory occurrences are selected and emphasized for perception.

Many phenomena related to attention have been described already by the ancient and medieval philosophers such as Aristotle, Lucretius, and Descartes. Into the scientific disciplines that preceded the modern psychology, attention was introduced in the 1740s by Wolff [1], and one of his observations was that the greater the attention, the smaller part of the visual field to which it extends.

Over the years, also many different terminologies and descriptions have been used [2], but at the moment there seems to be general agreement on at least the following fundamental features: 1. Attention is usually triggered by the onset of some external alerting occurrence (although it can also be evoked by various cues in mental tasks), 2. After the initial phase, attention is usually actively directed and narrowed, 3. Attention may be shifted involuntarily (e.g. by distractors), 4. The maintenance of attention for extended periods is difficult and easily interfered, 5. Cognitive interpretation of a prior stimulus has a biasing or *priming* effect on the direction of attention.

The objective of the work in presentation is very concrete: to give, by means of very simple modeling approaches, an explicit explanation to some of the most salient phenomena especially in visual attention, namely: 1. Activation of a part of the visual field during the fixation of the gaze, whereupon the width of the activated part will only depend on the "information content" of that part at a given relative resolution, 2. Narrowing of the activated field at more concentrated watching, whereupon "concentration" only means small eye movements, voluntary or involuntary.

In other words, I believe that at least the above two phenomena may have a very simple mechanistic interpretation, without any "consciousness" or "will" yet involved with them. Contrary to that, acts like involuntary shifting, priming, and even coordination of the muscular actions during attention may involve integrated cognitive and sensorimotor mechanisms and fall outside the scope of this discussion.

Modeling assumptions

I shall start with a few basic assumptions for modeling. These assumptions about the (mammalian) visual system will first be expressed in a general but yet concrete form in order to allow for various explicit implementations.

(i) The visual signals are propagated from the retina to the visual cortex via the thalamus along pathways that preserve their topological order all the way through.

(ii) The density of the retinal ganglia varies within wide limits over the retina. The signal pathways that start in the center of the retina are able to carry signals that represent denser variations in the spatial domain, while the peripheral pathways are better fit to smoother variations in the spatial domain.

(iii) With the main signal pathways there are associated other functions that modulate the transmittance of the former to the signals. These functions consist of neural circuits, each one of them monitoring and controlling its own subset of signal pathways up to a certain radius in the lateral direction. The associated circuits analyze the local "information content" of the signals, e.g., both spatial and temporal variations of the signals in their neighborhood, at different resolution. The control circuits also *compete* mutually in the sense that preference for the transmission of signals is given to those pathways, the associated control circuit of which detects the greatest variations, in relation to its own resolution, in the signals in its neighborhood. The rest of the neighboring pathways is inhibited or somehow neglected.

(iv) If the gaze is directed at a small object or densely located details concentrated on a small area, and there are no large optic structures in its surround, the control circuits associated with the narrow central areas of the retina are activated stronger than the peripheral control circuits. Accordingly, the central pathways will then be activated for signal transmission. Conversely, the broader or smoother the optic structures of the image around the gaze, the broader signal pathways are activated.

(v) As the control circuits are assumed sensitive to temporal changes in the spatial signal patterns, too, then, if the image is translated by a small amount on the retina, the relative changes caused by this translation are much bigger in the control circuits that modulate signals in the central area, where the resolution is high, than in the large peripheral areas with low resolution, respectively.

(vi) Thus, when the gaze is shifted by a small amount, activation of the central pathways in relation to the peripheral ones is increased, which is experienced as narrowing of the visual field and concentration of attention.

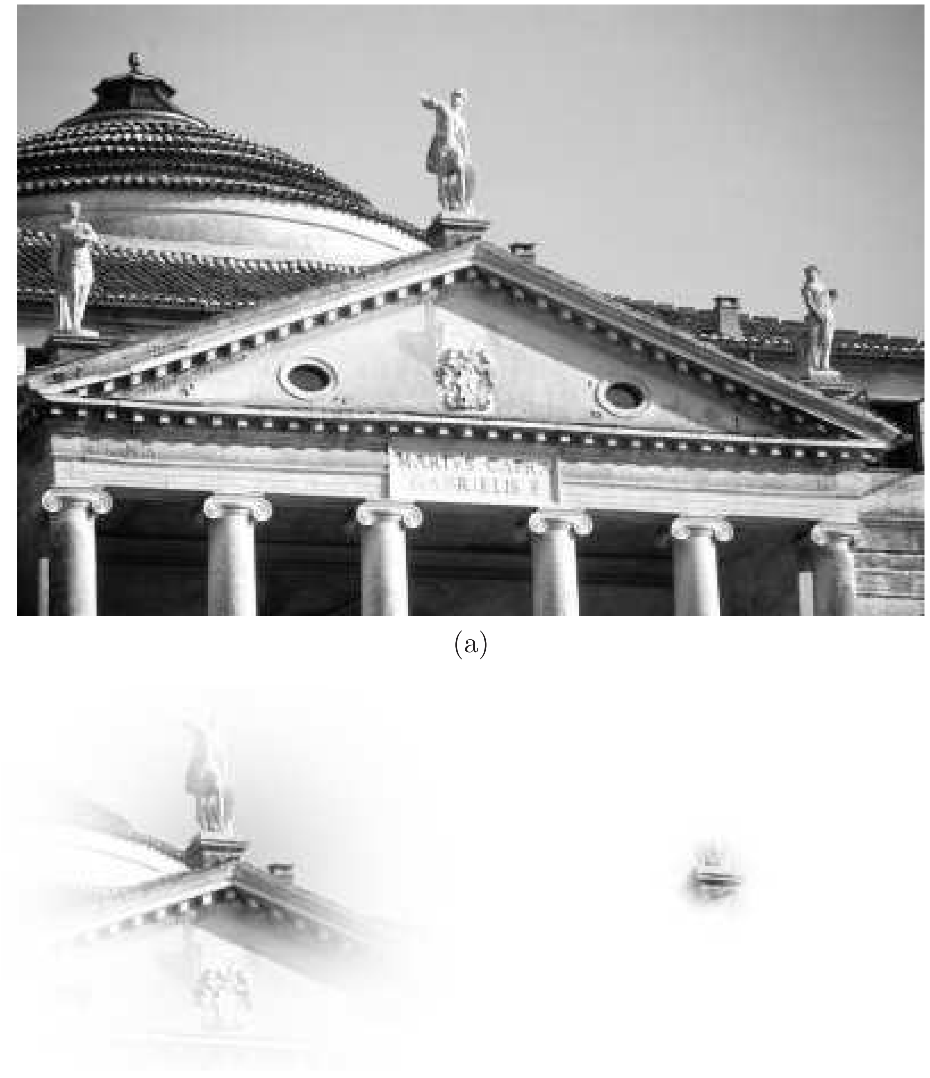
Simulations

In computer simulations I have used photographic images. The control circuits were assumed to lie symmetrically around the assumed direction of the gaze. The image intensities of these parts were modulated by the control circuits.

In a more extensive simulation one might want to place control circuits and pathways all over the surface of the simulated "retina" and also let them overlap. In this simulation, for simplicity, I have only used four overlapping and competing control circuits that lay concentrically around the direction of the gaze, with their widths relating approximately as 1:2:4:8.

Every control circuit shall in the first place detect whether there occur local spatial variations in the image at low resolution, and the control circuit shall be selective to optic structures the dimensions of which are of the same order of magnitude as the diameter of the circuit itself. Such an analysis can be performed in several ways. One possibility is that each control circuit operates like a low-level feature filter that may be implemented by laterally interconnected interneurons. In the present simulations, a functionally defined, very effective and computationally light method was used. A small number of local averages of the signals over each hypothetical control circuit were computed, and then the *variance* of these local averages was evaluated. The pathway over which the (relative) variance was biggest was activated for transmission.

In order to simulate the "narrowing" of the visual field under concentrated attention, the control circuits were also made sensitive to temporal variations in the signals. In the simplest way this was done in simulation by subtracting from each local average a fraction of the average computed previously, and evaluating the variances from these temporal differences of the spatial patterns.



(a)

(b) (c)

Figure 18.4: (a) The original image. (b) A part of the image was selected and activated by the "winning" control circuit. (c) When the gaze was moved by a small amount towards a detail (pedestal of the statue), the activated part of the visual field was further narrowed.

References

[1] Wolff, C. *Psychologia rationalis.* (Officina Libraria Rengeriana, Frankfurt & Leipzig 1740).

[2] Wright, R.D. (ed.) *Visual attention* (Oxford University Press, New York & Oxford 1998).

Publications of the Neural Networks Research Centre

Publications are in alphabetical order by the first author.

2000

1. Alhoniemi, E. Analysis of Pulping Data Using the Self-Organing Map. *Tappi Journal*, 2000. Vol. 83, No. 7, p. 66. http://www.tappi.org/public/tappi_journal.asp.

2. Bergius, P., Kiviluoto, K. & Maaranen, J. Cleaning Data with Self-Organizing Map to Improve Bankruptcy Risk Assesment. *Proc. of Advanced Investment Technologies'99 Conference*, Gold Coast, Australia, December 20-21, 1999 (2000). pp. 49-52.

3. Bingham, E. & Hyvärinen, A. A Fast Fixed-Point Algorithm for Independent Component Analysis of Complex Valued Signals. *International Journal of Neural Systems*, 2000. Vol. 10, No. 1, pp. 1-8.

4. Bingham, E. & Hyvärinen, A. Fast and Robust Deflationary Separation of Complex Valued Signals. *Proc. of X European Signal Processing Conference (EUSIPCO 2000)*, Tampere, Finland, September 4-8, 2000. pp. 23-26.

5. Bingham, E. & Hyvärinen, A. ICA of Complex Valued Signals: A Fast and Robust Deflationary Algorithm. *Proc. of International Joint Conference on Neural Networks (IJCNN 2000)*, Como, Italy, July 24-27, 2000. pp. 357-362.

6. Brandt, S., Laaksonen, J. & Oja, E. Statistical Shape Features in Content-Based Image Retrieval. *Proc. of 15th International Conference on Pattern Recognition*, Barcelona, Spain, September 3-7, 2000. Vol. 2, pp. 1066-1069.

7. Cristescu, R., Joutsensalo, J., Karhunen, J. & Oja, E. A Complexity Minimization Approach for Estimating Fading Gaussian Channel in CDMA Communications. *Proc. of 2nd International Workshop on Independent Component Analysis and Blind Signal Separation (ICA 2000)*, Helsinki, Finland, June 19-22, 2000. pp. 527-532.

8. Cristescu, R., Joutsensalo, J. & Ristaniemi, T. Fading Channel Estimation by Mutual Information Minimization for Gaussian Stochastic Processes. *Proc. of IEEE ICC2000, International Conference on Communications*, New Orleans, Louisiana, USA, June 18-22, 2000. pp. 56-59.

9. Cristescu, R., Karhunen, J., Joutsensalo, J. & Ristaniemi, T. Blind Separation Methods for CDMA Communications. *Proc. of Workshop on Machines That Learn*, Snowbird, Utah, April 4-7, 2000.

10. Cristescu, R., Ristaniemi, T., Joutsensalo, J. & Karhunen, J. Blind Separation of Convolved Mixtures for CDMA Systems. *Proc. of X European Signal Processing Conference (EUSIPCO 2000)*, Tampere, Finland, September 5-8, 2000. pp. 619-622.

11. Cristescu, R., Ristaniemi, T., Joutsensalo, J. & Karhunen, J. CDMA Delay Estimation Using Fast ICA Algorithm. *Proc. of IEEE International Symposium on Personal, Indoor, and Mobile Communications (PIMRC'00)*, London, UK, September 17-19, 2000. pp. 1117-1120.

12. Cristescu, R., Ristaniemi, T., Joutsensalo, J. & Karhunen, J. Delay Estimation in CDMA Communications Using A FastICA Algorithm. *Proc. of Second International Workshop on Independent Component Analysis and Blind Signal Separation (ICA 2000)*, Helsinki, Finland, June 19-22, 2000. pp. 105-110.

13. Cser, L., Mäntylä, P., Simula, O. & Ruha, P. Quality Control of Hot Rolling Mills Based on Data Mining and Knowledge Discovery. *Proc. of International ICSC Congress on Intelligent Systems & Applications*, Wollongong, Australia, December 11-15, 2000.

14. Cser, L., Simula, O., Mäntylä, P., Ruha, P. & Larkiola, J. Self-Organising Maps (SOM) in Hot Rolling. *Proc. of Intelligent Computation in Manufacturing Engineering (ICME 2000)*, Capri, Italy, June 21-23, 2000. pp. 291-297.

15. Federley, M., Alhoniemi, E., Laitila, M., Suojärvi, M. & Ritala, R. State Management for Process Monitoring, Diagnostics and Optimization. *Proc. of Control Systems 2000*, Victoria, British Columbia, Canada, May 1-4, 2000. pp. 295-298.

16. Flanagan, J. A. Neuron Weight Dynamics in the SOM and Self-Organized Criticality. *Proc. of International Joint Conference on Neural Networks (IJCNN 2000)*, Como, Italy, July 24-27, 2000. Vol. V, pp. 39-44.

17. Flanagan, J. A. Self-Organisation in the SOM with a Finite Number of Inputs. *Proc. of 8th European Symposium on Artificial Neural Networks (ESANN'2000)*, Bruges, Belgium, April 26-28, 2000. pp. 261-266.

18. Flanagan, J. A. Self-Organization in the SOM and Lebesque Continuity of the Input Distribution. *Proc. of International Joint Conference on Neural Networks (IJCNN 2000)*, Como, Italy, July 24-27, 2000. Vol. VI, pp. 26-34.

19. Flanagan, J. A. The Organized State in the SOM. *Proc. of STeP'2000, Finnish Artificial Intelligence Conference, 'AI of Tomorrow': Symposium on Theory*, Espoo, Finland, August 28-31, 2000. pp. 189-194.

20. Himberg, J. A SOM Based Cluster Visualization and Its Application for False Coloring. *Proc. of International Joint Conference on Neural Networks (IJCNN 2000)*, Como, Italy, July 24-27, 2000. pp. 581-586.

21. Hollmén, J., Skubacz, M. & Taniguchi, M. Input-Dependent Misclassification Costs for Cost-Sensitive Classification. *Proc. of Second International Conference on Data Mining 2000*, Cambridge, England, July 5-7, 2000. 2000, WIT Press, pp. 495-503.

22. Hollmén, J. & Tresp, V. Hidden Markov Model for Metric and Event-Based Data. *Proc. of 10th European Signal Processing Conference, EUSIPCO 2000*, Tampere, Finland, September 4-8, 2000. pp. 737-740.

23. Hollmén, J., Tresp, V. & Simula, O. Learning Vector Quantization Algorithm for Probabilistic Models. *Proc. of 10th European Signal Processing Conference, EUSIPCO 2000*, Tampere, Finland, September 4-8, 2000. pp. 721-724.

24. Honkela, A. & Valpola, H. Nonlinear Factor Analysis Matlab Package. Computer program. 2000. http://www.cis.hut.fi/projects/ica/bayes.

25. Hoyer, P. O. & Hyvärinen. A. Independent Component Analysis Applied to Feature Extraction from Colour and Stereo Images. *Network: Computation in Neural Systems*, 2000. Vol. 11, No. 3, pp. 191-210.

26. Hoyer, P. O. & Hyvärinen, A. Feature Extraction from Colour and Stereo Images Using ICA. *Proc. of International Joint Conference on Neural Networks (IJCNN 2000)*, Como, Italy, July 24-27, 2000. pp. 369-374.

27. Hoyer, P. O. & Hyvärinen, A. ICA Features of Colour and Stereo Images. *Proc. of International Workshop on Independent Component Analysis and Blind Signal Separation (ICA 2000)*, Helsinki, Finland, June 9-12, 2000. pp. 567-572.

28. Hoyer, P. O. & Hyvärinen, A. Modelling Chromatic and Binocular Properties of V1 Topography Using Topographic ICA. *Proc. of International Conference on Neural Information Processing (ICONIP'00)*, Taejon, Korea, November 14-18, 2000. pp. 1339-1344.

29. Hyvärinen, A. Complexity Pursuit: Combining Nongaussianity and Autocorrelations for Signal Separation. *Proc. of International Workshop on Independent Component Analysis and Blind Signal Separation (ICA 2000)*, Helsinki, Finland, June 19-22, 2000. pp. 175-180.

30. Hyvärinen, A. Complexity Pursuit: Using Both Nongaussianity and Autocorrelations for Signal Separation. *Proc. of International Conference on Neural Information Processing (ICONIP'00)*, Taejon, Korea, Nov. 14-18, 2000. pp. 1198-1203.

31. Hyvärinen, A. & Hoyer, P. O. Emergence of Phase and Shift Invariant Features by Decomposition of Natural Images into Independent Feature Subspaces. *Neural Computation*, 2000. Vol. 12, No. 7, pp. 1705-1720.

32. Hyvärinen, A. & Hoyer, P. O. Emergence of Topography and Complex Cell Properties from Natural Images Using Extensions of ICA. *Advances in Neural Information Processing Systems 12, Proc. of NIPS'99*, Denver, Colorado, USA, Nov. 30 - Dec. 2, 1999. 2000, MIT Press, pp. 827-833.

33. Hyvärinen, A. & Hoyer, P. O. Topographic ICA as a Model of V1 Receptive Fields and Topography. *Proc. of International Conference on Neural Information Processing (ICONIP'00)*, Taejon, Korea, Nov. 14-18, 2000. pp. 1323-1328.

34. Hyvärinen, A., Hoyer, P. O. & Inki, M. The Independence Assumption: Analyzing the Independence of the Components by Topography. In: Girolami, M. (ed.), *Advances in Independent Component Analysis.* 2000, Springer-Verlag, pp. 45-62.

35. Hyvärinen, A., Hoyer, P. O. & Inki, M. Topographic Independent Component Analysis: Visualizing the Independence Structure. *Proc. of International Workshop on Independent Component Analysis and Blind Signal Separation (ICA 2000)*, Helsinki, Finland, June 19-22, 2000. pp. 591-596.

36. Hyvärinen, A., Hoyer, P. & Inki, M. Topographic ICA as a Model of Natural Image Statistics. *Proc. of IEEE International Workshop on Biologically Motivated Computer Vision (BMCV'00)*, Seoul, Korea, May 15-17, 2000. pp. 535-544.

37. Hyvärinen, A., Hoyer, P. & Inki, M. Topographic ICA as a Model of V1 Receptive Fields. *Proc. of International Joint Conference on Neural Networks (IJCNN 2000)*, Como, Italy, July 24-27, 2000. pp. 83-88.

38. Hyvärinen, A. & Raju, K. Sparse Priors on the Mixing Matrix in Independent Component Analysis. *Proc. of International Workshop on Independent Component Analysis and Blind Signal Separation (ICA 2000)*, Helsinki, Finland, June 19-22, 2000. pp. 477-452.

39. Hyvärinen, A. & Oja, E. Independent Component Analysis: Algorithms and Applications. *Neural Networks*, 2000. Vol. 13, No. 4-5, pp. 411-430.

40. Iivarinen, J. Surface Defect Detection with Histogram-Based Texture Features. Intelligent Robots and Computer Vision XIX: Algorithms, Techniques, and Active Vision, *Proc. SPIE* 4197, 2000. pp. 140-145.

41. Iivarinen, J. Unsupervised Segmentation of Surface Defects with Simple Texture Measures. In: Pietikäinen, M. (ed.), *Texture Analysis in Machine Vision, Series in Machine Perception and Artificial Intelligence* - Vol. 40. 2000, World Scientific, pp. 231-238.

42. Iivarinen, J., Heikkinen, K., Rauhamaa, J., Vuorimaa, P. & Visa, A. A Defect Detection Scheme for Web Surface Inspection. *International Journal of Pattern Recognition and Artificial Intelligence*, 2000. Vol. 14, No. 6, pp. 735-755.

43. Karhunen, J., Malaroiu, S. & Ilmoniemi, M. Local Linear Independent Component Analysis Based on Clustering. *International Journal of Neural Systems*, 2000. Vol. 10, No. 6, pp. 439-451.

44. Kaski, S. Convergence of a Stochastic Semisupervised Clustering Algorithm. Espoo, Finland: 2000. (Helsinki University of Technology, Publications in Computer and Information Science Report A62).

45. Kaski, S. Kohti uutta luovan ajattelun opettamista. In: Hein, I. & Lauhia, R. (eds.), *OPE2, Dokumentoitua opetuksen kehittämistä Teknillisessä korkeakoulussa 1999-2000*. Helsinki, Finland 2000, Teknillinen korkeakoulu, Opetuksen ja opiskelun tuen julkaisuja 3/2000, pp. 9-14.

46. Kaski, S., Nikkilä, J. & Kohonen, T. Methods for Exploratory Cluster Analysis. *Proc. of SSGRR 2000, International Conference on Advances in Infrastructure for Electronic Business, Science, and Education on the Internet*, L'Aquila, Italy, July 31-August 6, 2000. (CD-ROM)

47. Kaski, S. & Sinkkonen, J. Metrics That Learn Relevance. *Proc. of International Joint Conference on Neural Networks (IJCNN 2000)*, Como, Italy, July 24-27, 2000. pp. 547-552.

48. Kaski, S., Venna, J. & Kohonen, T. Coloring that Reveals Cluster Structures in Multivariate Data. *Australian Journal of Intelligent Information Processing Systems*, 2000. Vol. 6, No. 2, pp. 82-88.

49. Kiviluoto, K., Bergius, P. & Maaranen, J. Business Applications of Neural Networks. In: Lisboa, P., Edisbury, B. & Vellido, A. (eds.), *Neural Networks for Analysis of Financial Statements*. Singapore 2000, World Scientific, pp. 73-84.

50. Kohonen, T. A Look into the Self-Organizing Maps. *Proc. of 2nd International ICSC Symposium on Neural Computation (NC'2000)*, Berlin, Germany, May 23-26, 2000. (CD-ROM)

51. Kohonen, T. An Unsupervised Learning Method that Produces Organized Representation from Real Information. *Proc. of ANNIMAB-1 Conference*, Göteborg, Sweden, May 13-16, 2000. London 2000, Springer, pp. 45-53.

52. Kohonen, T. New Lines in the Study of Self-Organizing Maps . *Proc. of 6th International Conference on Soft Computing, IIZUKA2000*, Iizuka, Japan, October 1-4, 2000. (CD-ROM)

53. Kohonen, T. Self-Organizing Maps of Massive Document Collections. *Proc. of International Joint Conference on Neural Networks (IJCNN'2000)*, Como, Italy, July 24-27, 2000. pp. 3-9.

54. Kohonen, T., Kaski, S., Lagus, K., Salojärvi, J., Honkela, J., Paatero, V. & Saarela, A. Self Organization of a Massive Document Collection. *IEEE Transactions on Neural Networks*, 2000. Vol. 11, pp. 574-585.

55. Koskela, M., Laaksonen, J., Laakso, S. & Oja, E. Content-Based Image Indexing and Retrieval: The PicSOM System. *Proc. of 9th Finnish Artificial Intelligence Conference (STeP 2000), 'AI of Tomorrow': Symposium on Theory*, Espoo, Finland. August 28-31, 2000. pp. 219-226.

56. Koskela, M., Laaksonen, J., Laakso, S. & Oja, E. Evaluating the Performance of Content-Based Image Retrieval Systems. *Proc. of Fourth International Conference on Visual Information Systems (VISual 2000)*, Lyon, France, November 2-4, 2000. pp. 430-441.

57. Koskela, M., Laaksonen, J., Laakso, S. & Oja, E. The PicSOM Retrieval System: Description and Evaluations. *Proc. of the Challenge of Image Retrieval 2000 (CIR2000)*, Brighton, UK, May 4-5, 2000. http://www.ewic.org.uk/ewic/

58. Kurimo, M. Indexing Spoken Audio by LSA and SOMs. *Proc. of 10th European Signal Processing Conference (EUSIPCO 2000)*, Tampere, Finland, September 4-8, 2000. pp. 2177-2180.

59. Laaksonen, J. T., Koskela, J. M., Laakso, S. P. & Oja, E. PicSOM - Content-Based Image Retrieval with Self-Organizing Maps. *Pattern Recognition Letters*, 2000. Vol. 21, No. 13-14, pp. 1199-1207.

60. Laaksonen, J., Oja, E., Koskela, M. & Brandt, S. Analyzing Low-Level Visual Features Using Content-Based Image Retrieval. *Proc. of 7th International Conference on Neural Information Processing (ICONIP'00)*, Taejon, Korea, November 14-18, 2000. pp. 1333-1338.

61. Lagus, K. Text Retrieval Using Self-Organized Document Maps. Espoo, Finland: 2000. 8 p. (Helsinki University of Technology, Publications in Computer and Information Science Report A61).

62. Laine, S., Pulkkinen, K. & Jämsä-Jounela, S.-L. On-Line Determination of the Concentrator Feed Type at Outokumpu Hitura Mine. *Minerals Engineering*, 2000. Vol. 13, No. 8-9, pp. 881-895.

63. Lappalainen, H. & Honkela, A. Bayesian Nonlinear Independent Component Analysis by Multi-Layer Perceptrons. In: Girolami, M. (ed.), *Advances in Independent Component Analysis*. 2000, Springer, pp. 93-121.

64. Lappalainen, H., Honkela, A., Giannakopoulos, X. & Karhunen, J. Nonlinear Source Separation Using Ensemble Learning and MLP Networks. *Proc. of Symposium 2000 on Adaptive Systems for Signal Processing, Communications, and Control (AS-SPCC)*, Lake Louise, Alberta, Canada, October 1-4, 2000. pp. 187-192.

65. Lappalainen, H. & Miskin, J. W. Ensemble Learning. In: Girolami, M. (ed.), *Advances in Independent Component Analysis*. 2000, Springer, pp. 76-92.

66. Lindblom, N., Heiskala, H., Hätönen, T., Mustanoja, S., Alfthan, H., Alila-Johansson, A. & Laakso, M.-L. No Evidence for Extraocular Light Induced Phase Shifting of Human Melatonin, Cortisol and Thyrotropin Rhythms. *Neuroreport*, 2000. Vol. 11, pp. 713-717.

67. Lindblom, N., Hätönen, T., Laakso, M.-L., Alila-Johansson, A., Laipio, M.-L. & Turpeinen, U. Bright Light Exposure of a Large Skin Area Does Not Affect Melatonin or Bilirubin Levels in Humans. *Biological Psychiatry*, 2000. Vol. 48, pp. 1098-1104.

68. Malaroiu, S., Kiviluoto, K. & Oja, E. ICA Preprocessing for Time Series Prediction. *Proc. of 2nd International Workshop on Independent Component Analysis and Blind Signal Separation (ICA 2000)*, Helsinki, Finland, June 19-22, 2000. pp. 453-457.

    Malaroiu, S., Kiviluoto, K. & Oja, E. Time Series Prediction with Independent Component Analysis. *Proc. of Advanced Investment Technologies'99 Conference*, Gold Coast, Australia, December 20-21, 1999 (2000). pp. 178-183.

69. Mustanoja, S.M., Hätönen, T., Alila-Johansson, A. & Laakso, M.-L. Evidence for Central α_2-Adrenergic Modulation of Rat Pineal Melatonin Synthesis. *Brain Research*, 2000. Vol. 887, pp. 174-177.

70. Oja, E., Kiviluoto, K. & Malaroiu, S. Independent Component Analysis for Financial Time Series. *Proc. of IEEE 2000 Symp. on Adapt. Systems for Signal Proc., Comm. and Control AS-SPCC*, Lake Louise, Canada, October 1-4, 2000. pp. 111-116.

71. Ojala, T., Valkealahti, K., Oja, E. & Pietikäinen, M. Texture Discrimination with Multidimensional Distributions of Signed Gray Level Differences. *Proc. of Asian Conference on Computer Vision (ACCV'2000)*, Taipei, Taiwan, Jan. 8-11, 2000. pp. 1082-1088.

72. Ollikainen, V., Bäckström, C. & Kaski, S. Automatic Content-based Sequential Ordering of Newspaper Articles for Electronic Books. *Proc. of International Conference on Engineering Applications of Neural Networks (EANN 2000)*, Kingston, UK, July 17-19, 2000. Kingston Upon Thames, England 2000, Kingston University, pp. 183-190.

73. Pajunen, P. & Girolami, M. Implementing Decisions in Binary Decision Trees Using Independent Component Analysis. *Proc. of Second International Workshop on Independent Component Analysis and Blind Signal Separation (ICA 2000)*, Helsinki, Finland, June 19-22, 2000. pp. 483-487.

74. Pajunen, P. & Karhunen, J. (eds.) *Proc. of Second International Workshop on Independent Component Analysis and Blind Signal Separation (ICA 2000)*, Helsinki, Finland, June 19-22, 2000.

75. Rantanen, J., Laine, S., Antikainen, O., Mannermaa, J-P., Simula, O. & Yliruusi, J. Visualization of Pharmaceutical Unit Operations Using Self-Organizing Maps (SOM). *J. Pharm. Biomed. Anal.*, 2000. Vol. 24, No. 3, pp. 343-352.

76. Siivola, V. Language Modeling Based on Neural Clustering of Words. Martigny, Switzerland: 2000. (IDIAP-Com Report 02).

77. Sinkkonen, J. & Kaski, S. Clustering by Similarity in an Auxiliary Space. *Proc. of Second International Conference on Intelligent Data Engineering and Automated Learning (IDEAL 2000)*, Hong Kong, December 13-15, 2000. Berlin 2000, Springer, pp. 3-8.

78. Sinkkonen, J. & Kaski, S. Semisupervised Clustering Based on Conditional Distributions in an Auxiliary Space. Espoo: 2000. 15 p. (Helsinki University of Technology, Publications in Computer and Information Science Report A60).

79. Sirola, M. Methodologies Used in Computerized Decision Support Systems of Safety Critical Processes - Comparison to Finnish Industry. *Proc. of IASTED International Conference for Modelling, Identification and Control*, Innsbruck, Austria, February 14-17, 2000. pp. 318-321.

80. Sirola, M. & Vesanto, J. Utilization of Neural Methods in Knowledge-Based Decision Support Systems - State Monitoring as a Case Example. *Proc. of IASTED International Conference for Modelling, Identification and Control*, Innsbruck, Austria, February 14-17, 2000. pp. 313-317.

81. Skubacz, M. & Hollmén, J. Quantization of Continuous Input Variables for Binary Classification. *Proc. of Second International Conference on Intelligent Data Engineering and Automated Learning (IDEAL'2000)*, Hong Kong, December 13-15, 2000. Berlin 2000, Springer-Verlag, pp. 42-47.

82. Somervuo, P. Competing Hidden Markov Models on the Self-Organizing Map. *Proc. of International Joint Conference on Neural Networks (IJCNN 2000)*, Como, Italy, July 24-27, 2000. vol. 3, pp. 169-174.

83. Somervuo, P. Speech Recognition Using Temporally Connected Kernels in Mixture Density Hidden Markov Models. *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2000)*, Istanbul, Turkey, June 5-9, 2000. pp. 3434-3437.

84. Somervuo, P. & Kohonen, T. Clustering and Visualization of Large Protein Sequence Databases by Means of an Extension of the Self-Organizing Map. Proc. of 3rd International Conference on Discovery Science, Kyoto, Japan, Dec. 4-6, 2000, *Lecture Notes in Artificial Intelligence* 1967. pp. 76-85.

85. Valpola, H. Nonlinear Independent Component Analysis Using Ensemble Learning: Theory. *Proc. of Second International Workshop on Independent Component Analysis and Blind Signal Separation (ICA 2000)*, Helsinki, Finland, June 19-22, 2000. pp. 251-256.

86. Valpola, H. Unsupervised Learning of Nonlinear Dynamic State-Space Models. Espoo, Finland: 2000. (Helsinki University of Technology, Publications in Computer and Information Science Report A59).

87. Valpola, H., Giannakopoulos, X., Honkela, A. & Karhunen, J. Nonlinear Independent Component Analysis Using Ensemble Learning: Experiments and Discussion. *Proc. of Second International Workshop on Independent Component Analysis and Blind Signal Separation (ICA 2000)*, Helsinki, Finland, June 19-22, 2000. pp. 351-356.

88. Valpola, H. & Pajunen, P. Fast Algorithms for Bayesian Independent Component Analysis. *Proc. of Second International Workshop on Independent Component Analysis and Blind Signal Separation (ICA 2000)*, Helsinki, Finland, June 19-22, 2000. pp. 233-237.

89. Venna, J. & Kaski, S. Coloring that Reveals Cluster Structures in Multidimensional Data. *Proc. of STeP'00, Finnish Artificial Intelligence Conference, 'AI of Tomorrow': Symposium on Theory*, Espoo, Finland, August 28-31, 2000. Helsinki, Finland 2000, Finnish Artificial Intelligence Society, pp. 203-210.

90. Vesanto, J. Neural Network Tool for Data Mining: SOM Toolbox. *Proc. of 5th International Symposium on Tool Environments and Development Methods for Intelligent Systems (TOOLMET2000)*, Oulu, Finland, April 13-14, 2000. Oulu, Finland 2000, Oulun yliopistopaino, pp. 184-196.

91. Vesanto, J. & Alhoniemi, E. Clustering of the Self-Organizing Map. *IEEE Transactions on Neural Networks*, 2000. Vol. 11 , No. 3, pp. 586-600.

92. Vesanto, J., Himberg, J. & Alhoniemi, E. SOM Toolbox for Matlab 5. Espoo, Finland: 2000. (Helsinki University of Technology, Publications in Computer and Information Science Report A57).

93. Vigário, R. Dipole Modeling in FastICA Decomposition of Evoked Responses. *Proc. of Second International Workshop on Independent Component Analysis and Blind Signal Separation (ICA 2000)*, Helsinki, Finland, June 19-22, 2000. pp. 363-368.

94. Vigário, R. & Oja, E. Independence: A New Criterion for the Analysis of the Electromagnetic Fields in the Global Brain? *Neural Networks*, 2000. Vol. 13, pp. 891-907.

95. Vigário, R., Särelä, J., Jousmäki, V., Hämäläinen, M. & Oja, E. Independent Component Approach to the Analysis of EEG and MEG Recordings. *IEEE Transactions on Biomedical Engineering*, 2000. Vol. 47, No. 5, pp. 589-593.

96. Vigário, R., Särelä, J. & Oja, E. Searching for Independence in Electromagnetic Brain Waves. In: Girolami, M. (ed.), *Advances in Independent Component Analysis.* London 2000, Springer, pp. 183-199.

97. Vuori, V., Aksela, M., Laaksonen, J., Oja, E. & Kangas, J. Adaptive Character Recognizer for a Hand-Held Device: Implementation and Evaluation Setup. *Proc. of 7th International Workshop on Frontiers in Handwriting Recognition*, Amsterdam, The Netherlands, Sep. 11-13, 2000. pp. 13-22.

98. Vuori, V., Laaksonen, J., Oja, E. & Kangas, J. Controlling On-Line Adaptation of a Prototype-Based Classifier for Handwritten Characters. *Proc. of 15th International Conference on Pattern Recognition (15 ICPR)*, Barcelona, Spain, September 3-7, 2000. pp. 331-334.

99. Vuori, V. & Oja, E. Analysis of Different Writing Styles with the Self-Organizin Map. *Proc. of 7th International Conference on Neural Information Processing (ICONIP'00)*, Taejon, Korea, November 14-18, 2000. pp. 1243-1247.

2001

1. Aksela, M., Laaksonen, J., Oja, E. & Kangas, J. Application of Adaptive Committee Classifiers in On-Line Character Recognition. *Proc. of Second International Conference on Advances in Pattern Recognition (ICAPR'01)*, Rio de Janeiro, Brazil, March 11-14, 2001. Berlin 2001, Springer, pp. 270-279.

2. Aksela, M., Laaksonen, J., Oja, E. & Kangas, J. Rejection Methods for an Adaptive Committee Classifier. *Proc. of Sixth International Conference on Document Analysis and Recognition (ICDAR'01)*, Seattle, Washington, USA, September 10-13, 2001. pp. 982-986.

3. Alhoniemi, E. & Simula, O. Interpretation and Comparison of Multidimensional Data Partitions. *Proc. of 9th European Symposium on Artificial Neural Networks (ESANN 2001)*, Bruges, Belgium, April 25-27, 2001. 2001, D-Facto, pp. 277-282.

4. Alila-Johansson, A., Eriksson, L., Soveri, T. & Laakso, M.-L. Seasonal Variation in Endogeneous Serum Melatonin Profiles in Goats: A Difference Between Spring and Fall? *Journal of Biological Rhythms*, 2001. Vol. 16, No. 3, pp. 254-263.

5. Bingham, E. Topic Identification in Dynamical Text by Extracting Minimum Complexity Time Components. *Proc. of 3rd International Conference on Independent Component Analysis and Blind Signal Separation (ICA 2001)*, San Diego, California, USA, Dec. 9-12, 2001. pp. 546-551.

6. Bingham, E., Kaban, A. & Girolami, M. Finding Topics in Dynamical Text: Application to Chat Line Discussions. *10th International World Wide Web Conference (WWW10)*, Hong Kong, May 1-5, 2001. Poster Proceedings, pp. 198-199.

7. Bingham, E. & Mannila, H. Random Projection in Dimensionality Reduction: Applications to Image and Text Data. *Proc. of 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2001)*, San Francisco, CA, USA, August 26-29, 2001. pp. 245-250.

8. Cser, L., Simula, O., Ruha, P. & Árvai, L. Quality Prediction based on Knowledge Extraction in Hot Rolling. *Proc. of 3rd International Conference on Intelligent Processing and Manufacturing of Materials (IPMM 2001)*, Vancouver, Canada, July 29-August 3, 2001. 6 p.

9. Cser, L., Simula, O., Ruha, P. & Árvai, L. Quality Prediction Based on Knowledge Extraction in Hot Rolling. *Proc. of International NAISO Symposium on Information Science Innovations in Intelligent Quality Management and Metrology (IQMM'2001)*, Dubai, U.A.E., March 17-21, 2001.

10. Flanagan, J. A. Self-Organization in the One-Dimensional SOM with a Decreasing Neighborhood. *Neural Networks*, 2001. Vol. 14, pp. 1405-1417.

11. Flanagan, J. A. Self-Organized Criticality and the Self-Organizing Map. *Physical Review E*, 2001. Vol. 63, p. 36130-1-5.

12. Flanagan, J. A. Topologically Ordered Neural Networks. In: Moss, F. & Gielen, S. (eds.), *Neuro-Informatics and Neural Modelling, Handbook of Biological Physics*, Vol. 4. Amsterdam 2001, Elsevier, pp. 685-729.

13. Funaro, M., Oja, E. & Valpola, H. Artefact Detection in Astrophysical Image Data Using Independent Component Analysis. *Proc. of 3rd International Conference on Independent Component Analysis and Blind Signal Separation (ICA 2001)*, San Diego, California, USA, Dec. 9-12, 2001. pp. 43-48.

14. Giannakopoulos, X. & Valpola, H. Nonlinear Dynamical Factor Analysis. *Proc. of Twentieth International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, MaxEnt 2000*, Paris, France, July 8-13, 2000. Melville, NY 2001, American Institute of Physics, pp. 305-317.

15. Girdziusas, R. & Laaksonen, J. Multilayer Perceptron Approach to Nonrigid Image Matching. *Proc. of 8th International Conference on Neural Information Processing (ICONIP 2001)*, Shanghai, China, Nov. 14-18, 2001. pp. 491-496.

16. Girolami, M. A Variational Method for Learning Sparse and Overcomplete Representations. *Neural Computation*, 2001. Vol. 13, No. 11, pp. 2517-2532.

17. Himberg, J., Ahola, J., Alhoniemi, E., Vesanto, J. & Simula, O. The Self-Organizing Map as a Tool in Knowledge Engineering. In: Pal, Nikhil R. (ed.), *Pattern Recognition in Soft Computing Paradigm, Volume 2 of FLSI Soft Computing Series*. 2001, World Scientific, pp. 38-65.

18. Himberg, J. & Hyvärinen, A. Independent Component Analysis for Binary Data: An Experimental Study. *Proc. of 3rd International Conference on Independent Component Analysis and Blind Signal Separation (ICA2001)*, San Diego, California, USA, Dec. 9-12, 2001. pp. 552-556.

19. Honkela, A. & Karhunen, J. An Ensemble Learning Approach to Nonlinear Independent Component Analysis. *Proc. of European Conference on Circuit Theory and Design (ECCTD'01)*, Espoo, Finland, August 28-31, 2001. Vol. 1, pp. 41-44.

20. Hoyer, P. O. & Hyvärinen, A. Emergence of Contour Coding and Integration in a Sparse Coding Neural Network. *Proc. of Internatioal Conference on Cognitive and Neural Systems (ICCNS)*, Boston, MA, USA, May 31 - June 2, 2001.

21. Hyvärinen, A. Blind Source Separation by Nonstationarity of Variance: A Cumulant-Based Approach. *IEEE Trans. on Neural Networks*, 2001. Vol. 12, No. 6, pp. 1471-1474.

22. Hyvärinen, A. Complexity Pursuit: Separating Interesting Components from Time-Series. *Neural Computation*, 2001. Vol. 13, No. 4, pp. 883–898.

23. Hyvärinen, A. Fast Independent Component Analysis. In: Roberts, S. & Everson, R. (eds.), *ICA: Principles and Practice*. Cambridge, UK 2001, Cambridge University Press, pp. 71-94.

24. Hyvärinen, A. & Hoyer, P. O. Topographic Independent Component Analysis as a Model of V1 Organization and Receptive Fields. *Neurocomputing*, 2001. Vol. 38-40, No. 1-4, pp. 1307-1315.

25. Hyvärinen, A. & Hoyer, P.O. A Two-Layer Sparse Coding Model Learns Simple and Complex Cell Receptive Fields and Topography from Natural Images. *Vision Research*, 2001. Vol. 41, No. 18, pp. 2413-2423.

26. Hyvärinen, A., Hoyer, P.O & Inki, M. Topographic Independent Component Analysis. *Neural Computation*, 2001. Vol. 13, No. 7, pp. 1527-1558.

27. Hyvärinen, A., Hoyer, P. & Oja, E. Image Denoising by Sparse Code Shrinkage. In: Haykin, S. & Kosko, B., *Intelligent Signal Processing.* New York 2001, IEEE Press, pp. 554-568.

28. Hyvärinen, A., Karhunen, J. & Oja, E. *Independent Component Analysis.* New York 2001, John Wiley.

29. Iivarinen, J. Defect Detection of Texture Surfaces. *Proc. of IASTED International Conference on Visualization, Imaging, and Image Processing*, Marbella, Spain, September 3-5, 2001. pp. 295-300.

30. Iivarinen, J. Unsupervised Segmentation of Defect Images. Intelligent Robots and Computer Vision XX: Algorithms, Techniques, and Active Vision, (Newton, Massach., Oct. 29-31, 2001), *Proc.S̃PIE 4572.* pp. 488-495.

31. Iline, A., Valpola, H. & Oja, E. Detecting Process State Changes by Nonlinear Blind Source Separation. *Proc. of 3rd International Conference on Independent Component Analysis and Blind Signal Separation (ICA 2001)*, San Diego, California, USA, Dec. 9-12, 2001. pp. 704-709.

32. Inki, M. & Hyvärinen, A. Two Methods for Estimating Overcomplete Independent Component Bases. *Proc. of 3rd International Conference on Independent Component Analysis and Blind Signal Separation (ICA 2001)*, San Diego, California, USA, December 9-12, 2001. pp. 343-348.

33. Joutsiniemi, S.-L., Nurminen, J. & Kohonen, T. SOM in Detecting States of Sleep and Wakefulness in Polysomnographic Data. *Proc. of Fourth International Conference on Neural Networks and Expert Systems in Medicine and Healthcare*, Milos Island, Greece, September 20-22, 2001. pp. 97-101.

34. Karhunen, J. Nonlinear Independent Component Analysis. In: Everson, R. & Roberts, S. (eds.), *ICA: Principles and Practice.* Cambridge, UK 2001, Cambridge University Press, pp. 113-134.

35. Kaski, S. Learning Metrics for Exploratory Data Analysis. *Proc. of 2001 IEEE Signal Processing Society Workshop, Neural Networks for Signal Processing XI*, Falmouth, Massach., USA, Sept. 10-12, 2001. New York, NY 2001, IEEE, pp. 53-62.

36. Kaski, S. SOM-Based Exploratory Analysis of Gene Expression Data. In: *Advances in Self-Organizing Maps*, Proc. of Workshop on Self-Organizing Maps (WSOM'01), Lincoln, UK, June 13-15, 2001. London 2001, Springer, pp. 124-131.

37. Kaski, S., Nikkilä, J., Törrönen, P., Castren, E. & Wong, G. Analysis and Visualization of Gene Expression Data Using Self-Organizing Maps. *Proc. of IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing (NSIP-01)*, Baltimore, Maryland, USA, June 3-6, 2001.

38. Kaski, S. & Salojärvi, J. Generative Mixture Modeling by Autonomous Estimators. *Proc. of 5th International Conference Knowledge-Based Intelligent Information Engineering Systems & Allied Technologies (KES'2001)*, Osaka, Japan, Sept. 6-8, 2001. Amsterdam 2001, IOS Press, pp. 250-254.

39. Kaski, S. & Sinkkonen, J. A Topography-Preserving Latent Variable Model with Learning Metrics. In: *Advances in Self-Organizing Maps*, Proc. of Workshop on Self-Organizing Maps (WSOM'01), Lincoln, UK, June 13-15, 2001. London 2001, Springer, pp. 224-229.

40. Kaski, S., Sinkkonen, J. & Nikkilä, J. Clustering Gene Expression Data by Mutual Information with Gene Function. *Proc. of International Conference on Artificial Neural Networks (ICANN 2001)*, Vienna, Austria, August 21-25, 2001. Berlin 2001, Springer, pp. 81-86.

41. Kaski, S., Sinkkonen, J. & Peltonen, J. Bankruptcy Analysis with Self-Organizing Maps in Learning Metrics. *IEEE Transactions on Neural Networks*, 2001. Vol. 12, pp. 936-947.

42. Kaski, S., Sinkkonen, J. & Peltonen, J. Data Visualization and Analysis with Self-Organizing Maps in Learning Metrics. *Proc. of Third International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2001)*, Munich, Germany, Sept. 5-7, 2001. Berlin 2001, Springer, pp. 162-173.

43. Kaski, S., Sinkkonen, J. & Peltonen, J. Learning Metrics for Self-Organizing Maps. *Proc. of International Joint Conference on Neural Networks (IJCNN'01)*, Washington, D.C., July 15-19, 2001. Piscataway, NJ 2001, IEEE, pp. 914-919.

44. Kohonen, T. A Neural Network That Emphasizes or Suppresses Visual Objects According to Eye Movements. Espoo, Finland: Helsinki University of Technology, 2001. 14 p. (Publications in Computer and Information Science, Report A64).

45. Kohonen, T. Overture. In: Seiffert, U. & Jain, L.C. (eds.), *Self-Organizing Neural Networks, Recent Advances and Applications*, Studies in Fuzziness and Soft Computing, Vol. 78. Heidelberg 2001, Springer-Verlag, pp. 1-12.

46. Kohonen, T. Perspectives on the Research of Artificial Neural Networks. *International Journal of Computer Research*, 2001. Vol. 10, No. 2, pp. 127-138.

47. Kohonen, T. *Self-Organizing Maps.* (3rd ed.). Heidelberg 2001, Springer-Verlag.

48. Kohonen, T. Self-Organizing Maps of Massive Databases. *Engineering Intelligent Systems*, 2001. Vol. 4, pp. 179-185.

49. Kohonen, T. Voivatko koneet oppia omaehtoisesti? In: Heino, P. & Westermarck, H. (eds.), *Studia Generalia 2000, Matematiikka, kulttuurimme perusta - ja trauma.* Helsinki 2001, Helsingin yliopiston vapaan sivistystyön toimikunta, pp. 81-92. (in Finnish)

50. Koivunen, V., Enescu, M. & Oja, E. Adaptive Algorithm for Blind Separation from Noisy Time-Varying Mixtures. *Neural Computation*, 2001. Vol. 13, No. 10, pp. 2339-2357.

51. Koskela, M., Laaksonen, J. & Oja, E. Comparison of Techniques for Content-Based Image Retrieval. *Proc. of 12th Scandinavian Conference on Image Analysis (SCIA'01)*, Bergen, Norway, June 11-14, 2001. pp. 579-586.

52. Koskela, M., Laaksonen, J. & Oja, E. PicSOM Content-Based Image Retrieval System - Comparison of Techniques. *Proc. of 3rd European Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2001)*, Tampere, Finland, May 16-17, 2001.

53. Koskela, M., Laaksonen, J. & Oja, E. Self-Organizing Image Retrieval with MPEG-7 Descriptors. *Proc. of Infotech Oulu International Workshop on Information Retrieval (IR 2001)*, Oulu, Finland, September 19-21, 2001.

54. Laakso, S., Laaksonen, J., Koskela, M. & Oja, E. Self-Organizing Maps of Web Link Information. In: *Advances in Self-Organizing Maps*, Proc. of Workshop on Self-Organizing Maps (WSOM'01), Lincoln, UK, June 13-15, 2001. London 2001, Springer, pp. 146-151.

55. Laaksonen, J., Koskela, M., Laakso, S. & Oja, E. Self-Organising Maps as a Relevance Feedback Technique in Content-Based Image Retrieval. *Pattern Analysis & Applications*, 2001. Vol. 4, pp. 140-152.

56. Lagus, K. Studying Similarities in Term Usage with Self-Organizing Maps. *Proc. of NordTerm'01*, Tuusula, Finland, June 13-16, 2001.

57. Lagus, K. & Airola, A. Analysis of Functional Similarities of Finnish Verbs Using the Self-Organizing Map. *Proc. of 13th European Summer School in Logic, Language and Information (ESSLLI'01)*, Helsinki, Finland, Aug. 13-24, 2001.

58. Lagus, K., Alhoniemi, E. & Valpola, H. Independent Variable Group Analysis. *Proc. of International Conference on Artificial Neural Networks (ICANN 2001)*, Vienna, Austria, August 21-25, 2001. Berlin 2001, Springer, pp. 203-210.

59. Laine, S. Combining Off-Line and On-Line Information in Process Study Using the Self- Organizing Map (SOM). *Proc. of IEEE Mountain Workshop on Soft Computing in Industrial Applications*, Blacksburg, Virginia, USA, June 25-27, 2001. pp. 71-76.

60. Laine, S. State Based Process Study Using the SOM and a Variable Selection Technique. *Proc. of International Conference on Engineering Applications of Neural Networks*, Cagliari, Italy, July 16-18, 2001. pp. 15-22.

61. Lindblom, N., Heiskala, H., Kaski, M., Leinonen, L., Nevanlinna A., Iivanainen, M. & Laakso, M.-L. Neurological Impairments and Sleep-Wake Behavior among the Mentally Retarded. *Journal of Sleep Research*, 2001. Vol. 10, pp. 309-318.

62. Luukka, P., Saastamoinen, K. & Könönen, V. A Classifier Based on the Maximal Fuzzy Similarity in the Lukasiewicz-Structure. *Proc. of 10th IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2001)*, Melbourne, Australia, Dec. 2-5, 2001.

63. Nikkilä, J., Törönen, P., Sinkkonen, J. & Kaski, S. Analysis of Gene Expression Data Using Semi-Supervised Clustering. *Proc. of Bioinformatics 2001*, Skövde, Sweden, March 29-April 1, 2001. pp. 2-3.

64. Oja, E. Machine Learning in Data Mining. *Proc. of 10th International PEG Conference (PEG 2001)*, Tampere, Finland, June 23-26, 2001.

65. Oja, E., Finding Structure in Signals, Images, and Data. *Proc. of 8th Int. Conf. on Neural Information Processing (ICONIP2001)*, Shanghai, China, Nov. 14-18, 2001. pp. 9-13.

66. Ojala, T., Valkealahti, K., Oja, E. & Pietikäinen, M. Texture Discrimination with Multidimensional Distributions of Signed Gray-Level Differences. *Pattern Recognition*, 2001. Vol. 34, No. 3, pp. 727-739.

67. Peura, M. Trix - A C++ Software Package for Pattern Recognition. Computer program. 2001. http://www.cis.hut.fi/peura/Trix/

68. Peura, M. & Syrjäsuo, M. Flexible Heuristic Matching of Attribute Trees. *Proc. of 12th Scandinavian Conference on Image Analysis*, Bergen, Norway, June 11-14, 2001. pp. 207-213.

69. Raiko, T. & Valpola, H. Missing Values in Nonlinear Factor Analysis. *Proc. of 8th International Conference on Neural Information Processing (ICONIP 2001)*, Shanghai, China, Nov. 14-18, 2001. pp. 822-827.

70. Raivio, K., Simula, O. & Laiho, J. Neural Analysis of Mobile Radio Access Network. *Proc. of 2001 IEEE International Conference on Data Mining (ICDM'01)*, San Jose, California, USA, Nov. 29 - Dec. 2, 2001. pp. 457-464 .

71. Sevon, P., Ollikainen, V., Onkamo, P., Toivonen, H.T.T., Mannila, H. & Kere, J. Mining Associations Between Genetic Markers, Phenotypes and Covariates. *Genetic Epidemiology*, 2001. Vol. 21 (Suppl. 1), pp. S588-S593.

72. Siivola, V., Kurimo, M. & Lagus, K. Large Vocabulary Statistical Language Modeling for Continuous Speech Recognition in Finnish. *Proc. of 7th European Conference on Speech Technology and Communication, EUROSPEECH 2001*, Aalborg, Denmark, Sept. 3-7, 2001. pp. 737-740.

73. Simula, O., Hollmén, J. & Alhoniemi, E. Models from Data: Analysis of Industrial Processes and Telecommunication Systems. *Proc. of Automazione'2001*, Ancona, Italy, Nov. 22-23, 2001. pp. 13-19.

74. Sinkkonen, J. & Kaski, S. Clustering Based on Conditional Distributions in an Auxiliary Space. *Neural Computation*, 2001. Vol. 14, pp. 217-239.

75. Siponen, M., Vesanto, J., Simula, O. & Vasara, P. An Approach to Automated Interpretation of SOM. In: *Advances in Self-Organizing Maps*, Proc. of Workshop on Self-Organizing Maps (WSOM'01), Lincoln, England, June 13-15, 2001. London 2001, Springer, pp. 89-94.

76. Sirola, M. Applying Decision Analysis Method in Process Control Problem in Accident Management Situation. *Proc. of 14th International Conference on Systems Science*, Wroclaw, Poland, Sept. 11-14, 2001. pp. 70-76.

77. Sirola, M. Decision Model for State Monitoring. *Proc. of IASTED International Conference for Modelling, Identification and Control*, Innsbruck, Austria, Feb. 19-22, 2001. pp. 350-353.

78. Sirola, M. Some Ideas of Using SOM Analysis in Industrial Applications. *Proc. of IASTED International Conference for Modelling, Identification and Control*, Innsbruck, Austria, Feb. 19-22, 2001. pp. 439-441.

79. Särelä, J., Valpola, H., Vigário, R. & Oja, E. Dynamical Factor Analysis of Rhythmic Magnetoencephalographic Activity. *Proc. of 3rd International Conference on Independent Component Analysis and Blind Signal Separation (ICA 2001)*, San Diego, California, USA, Dec. 9-12, 2001. pp. 451-456.

80. Särelä, J. & Vigário, R. The Problem of Overlearning in High-Order ICA Approaches: Analysis and Solutions. *Proc. of International Work-conference on Artificial and Natural Neural Networks (IWANN'01)*, Granada, Spain, June 13-15, 2001. pp. 818-825.

81. Valpola, H. Learning Hidden Dynamical Systems. *Proc. of The Learning Workshop*, Snowbird, Utah, USA, April 10-13, 2001. 2 p.

82. Valpola, H., Honkela, A. & Karhunen, J. Nonlinear Static and Dynamic Blind Source Separation Using Ensemble Learning. *Proc. of International Joint Conference on Neural Networks (IJCNN'01)*, Washington D.C., USA, July 15-19, 2001. pp. 2750-2755.

83. Valpola, H., Raiko, T. & Karhunen, J. Building Blocks for Hierarchical Latent Variable Models. *Proc. of 3rd International Conference on Independent Component Analysis and Blind Signal Separation (ICA 2001)*, San Diego, California, USA, Dec. 9-12, 2001. pp. 710-715.

84. Venna, J. & Kaski, S. Neighborhood Preservation in Nonlinear Projection Methods: An Experimental Study. *Proc. of International Conference on Artificial Neural Networks (ICANN 2001)*, Vienna, Austria, August 21-25, 2001. Berlin 2001, Springer, pp. 485-491.

85. Vesanto, J. Importance of Individual Variables in the k-means Algorithm. *Proc. of 5th Pacific-Area Conference on Knowledge Discovery in Databases (PAKDD2001)*, Hong Kong, April 16-18, 2001. pp. 513-518.

86. Vigário, R. From Principal to Independent Component Analysis of Brain Signals. *Proc. of 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'2001)*, Istanbul, Turkey, October 25-28, 2001. (CD-ROM)

87. Vuori, V., Laaksonen, J., Oja, E. & Kangas, J. Experiments with Adaptation Strategies for a Prototype-Based Recognition System of Isolated Handwritten Characters. *International Journal of Document Analysis and Recognition IJDAR*, 2001. Vol. 3, No. 3, pp. 150-159.

88. Vuori, V., Laaksonen, J., Oja, E. & Kangas, J. Speeding up On-line Recognition of Handwritten Characters by Pruning the Prototype Set. *Proc. of Sixth International Conference on Document Analysis and Recognition (ICDAR'01)*, Seattle, Washington, USA, September 10-13, 2001. pp. 501-505.

II From Data to Knowledge Research Unit
Research Projects under the CIS Laboratory

Chapter 19

From Data to Knowledge Research Unit

Heikki Mannila, Jaakko Hollmén, Jouni K. Seppänen, Kalle Korpiaho, Johanna Tikanmäki, Ella Bingham

19.1 Data mining: discrete pattern discovery and probabilistic techniques

The pattern discovery group works on different techniques for the discovery of discrete patterns and on methods for finding various types of probabilistic models for discrete data. An overview of some of the methods was published in 2001 [1].

Modeling of sequences: partial orders, components, and random projections

An example of the work is the article [2]. The problem considered is finding good descriptions of data sets consisting of sequences of discrete values. The chosen model class consists of partial orders: a partial order can be considered to describe all total orders (i.e., sequences) that are compatible with it. Thus the task is to find a collection of partial orders that describes as large a fraction of the sequences in the input data as possible. In addition to this requirement, the partial orders should be as specific as possible, i.e., the number of all total orders compatible with the partial orders should be small. Computing the number of all total orders that are compatible with a partial order is a $\sharp$P-complete problem, so the model class is restricted to the set of series-parallel partial orders.

These requirements specify an optimization problem of finding a set of partial orders maximizing a scoring function. In [2] an algorithm is given for this problem, and empirical results are reported.

Long sequences of data items arise in many applications, such as telecommunications network management. One of the problems there is locating whether a recent combination of events has appeared before. In [3] we considered the use of random projections in finding similar situations. It turns out that the basic random projection approach can be profitably applied also in this situation.

Another problem in the analysis of sequences is finding whether a sequence of discrete symbols can be considered to be the independent mixture of components. That is, the task is to find a discrete analogue of independent component analysis (ICA). In [4] we considered ways of defining this notion and gave an algorithm for this task.

A sequence of discrete events can be described by giving the intensity of the events, assuming that each event is generated by a Poisson process with piecewise constant intensity. In [5] we use Markov chain Monte Carlo to obtain posterior distributions on the intensity. While this method is flexible, it is computationally very intensive. We considered also approaches based on dynamic programming, and showed that careful pruning of the candidate space of change points of the intensity functions can be used to obtain very fast methods for finding nearly optimal descriptions. Similar techniques are used in time series segmentation for the purpose of context recognition in mobile phones in [6].

Large-dimensional discrete data

Various applications, such as market basket data analysis and document analysis, result in large collections of high-dimensional count data. Such data sets are typically sparse: most entries are zero. The basic fast counting technique for such data sets is the finding of frequent sets and association rules. While there are lots of algorithms for this task, actual use of the methods is rare, as the resulting summaries of the data are difficult to use.

We have considered different ways of using the frequent set to obtain more useful information. One interesting task is to find similarities between attributes on the basis of context information. The basic idea is that two 0-1 attributes are similar, if they occur in similar contexts. For example, in market basket data Coke and Pepsi can be considered to be similar, if the buying behavior of Coke buyers is similar to the buying behavior of Pepsi buyers. An iterative algorithm based on this idea is given in [7]. More detailed probabilistic modeling of market basked data is considered in [8].

Random projection in dimensionality reduction

In many applications of data mining, the high dimensionality of the data restricts the choice of data processing methods. A statistically optimal way of dimensionality reduction is to project the data onto a lower-dimensional orthogonal subspace that captures as much of the variation of the data as possible. The best (in mean-square sense) and most widely used way to do this is principal component analysis (PCA); unfortunately it is quite expensive to compute for high-dimensional data sets, as it requires the eigenvalue decomposition of the data covariance matrix.

A computationally simple method of dimensionality reduction that does not introduce a significant distortion in the data set would thus be desirable. Random projection (RP) has recently emerged as a promising alternative for this task. In random projection, the original d-dimensional data matrix X is projected onto a k-dimensional ($k \ll d$) subspace through the origin, using a random $k \times d$ matrix R whose columns have unit lengths. Using matrix notation where $X_{d\times N}$ is the original set of N d-dimensional observations,

$$X^{RP}_{k\times N} = R_{k\times d} X_{d\times N} \tag{19.1}$$

is the projection of the data onto a lower k-dimensional subspace. The key idea of random mapping arises from the Johnson-Lindenstrauss lemma [9]: if points in a vector space are projected onto a randomly selected subspace of suitably high dimension, then the distances between the points are approximately preserved.

While this method has attracted lots of interest, empirical results are sparse. For this reason, we have presented experimental results on using random projection as a dimensionality reduction tool in the context of both noisy and noiseless images, and information retrieval in text documents [10]. As an example, we present here results on monochrome images of natural scenes. The distortion caused by dimensionality reduction is measured by comparing the Euclidean distance between two dimensionality reduced data vectors to their Euclidean distance in the original high-dimensional space. Figure 19.1 shows the mean squared error in the distance between members of a pair of data vectors, averaged over 100 pairs. The results of random projection with a Gaussian distributed random matrix (RP), random projection with a sparse random matrix [11] (SRP), principal component analysis (PCA) and discrete cosine transform (DCT) are shown, together with their 95 per cent confidence intervals. Figure 19.2 shows the number of Matlab's floating point operations needed when using the abovementioned methods. It can be seen that PCA is significantly more burdensome than RP or DCT.

To conclude, we have shown that projecting the data onto a random lower-dimensional subspace yields results comparable to conventional dimensionality reduction methods such as principal component analysis: the similarity of data vectors is preserved well under random projection. We have also shown experimentally that using a sparse random matrix [11] gives additional computational savings in random projection.

References

[1] D. Hand, H. Mannila and P. Smyth. *Principles of Data Mining.* MIT Press 2001. ISBN 0-262-98290-X.

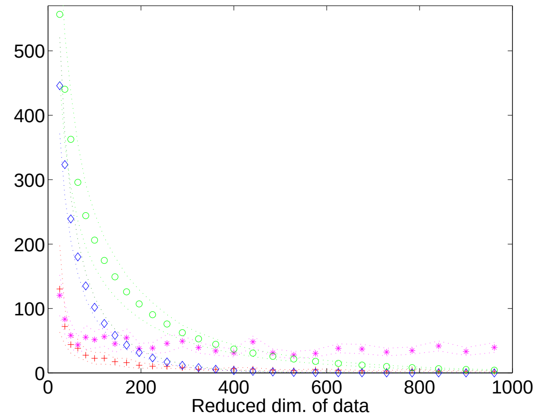


Figure 19.1: MSE of Euclidean distances, and 95 % confidence intervals over 100 pairs of image vectors. RP+ sparse RP$*$ PCA$\diamond$ DCT$\circ$

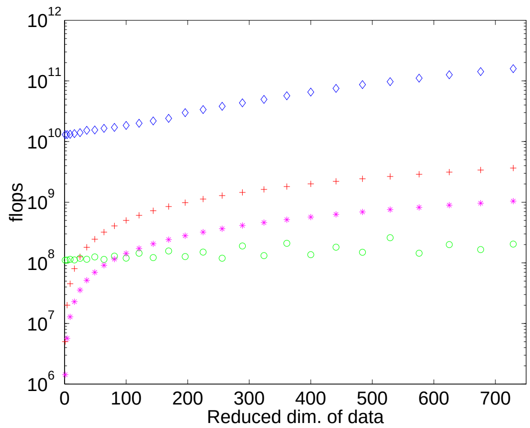


Figure 19.2: Number of Matlab's floating point operations needed when reducing the dimensionality of image data using RP+ sparse RP$*$ PCA$\diamond$ DCT$\circ$

[2] H. Mannila and C. Meek. Global partial orders from sequential data. *Sixth Annual Conference on Knowledge Discovery and Data Mining (KDD-2000)*, pp. 161–168.

[3] H. Mannila and J.K. Seppänen. Recognizing similar situations from event sequences. *First SIAM Conference on Data Mining*, 2001. `http://www.siam.org/meetings/sdm01/pdf/sdm01_03.pdf`

[4] H. Mannila and D. Rusakov. Decomposing event sequences into independent components. *First SIAM Conference on Data Mining*, 2001. `http://www.siam.org/meetings/sdm01/pdf/sdm01_02.pdf`

[5] H. Mannila and M. Salmenkivi. Finding simple intensity descriptions from event sequence data. *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2001)*, F. Provost and R. Srikant (eds.), pp. 341–346.

[6] J. Himberg, K. Korpiaho, H. Mannila, J. Tikanmäki and H. Toivonen. Time-series segmentation for context recognition in mobile devices. *IEEE Conference on Data Mining* 2001, pp. 203–207.

[7] G. Das and H. Mannila. Context-based similarity methods for categorical attributes. *Principles of Data Mining and Knowledge Discovery, 4th European Conference (PKDD 2000)*, D.A. Zighed et al. (eds.), pp. 201–211.

[8] I. Cadez, P. Smyth and H. Mannila. Probabilistic Modeling of Transaction Data with Applications to Profiling, Visualization, and Prediction. *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2001)*, F. Provost and R. Srikant (eds.), pp. 37–46.

[9] W.B. Johnson and J. Lindenstrauss. Extensions of Lipshitz mapping into Hilbert space. In *Conference in modern analysis and probability*, volume 26 of *Contemporary Mathematics*, pages 189–206. Amer. Math. Soc., 1984.

[10] E. Bingham and H. Mannila. Random projection in dimensionality reduction: applications to image and text data. In *Proc. 7th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD-2001)*, pages 245–250, 2001.

[11] D. Achlioptas. Database-friendly random projections. In *Proc. ACM Symp. on the Principles of Database Systems*, pages 274–281, 2001.

[12] C.K. Leung, R. Ng, and H. Mannila. Segmentation for frequency counting. *ICDE 2002*, to appear.

[13] H. Mannila. Theoretical frameworks for data mining. *SIGKDD Explorations* 1, 2 (January 2000), pp. 30–32

[14] J. Mäntyjärvi, J. Himberg, P. Korpipää, and H. Mannila. Extracting the context of a mobile device user. To appear.

19.2 Applications in bioinformatics

Gene Expression

The research of the Pattern Discovery group in genetic expression data has been concentrated in two main directions. First, we have examined publically available data on expression in baker's yeast in order to understand the fundaments of the field and to develop new techniques; second, we have applied proven techniques to data on expression in humans.

All biological organisms contain genetic material in the form of DNA molecules. These molecules are copied in the cell division process, and thus all cells of a single organism contain copies of the same DNA. The differences in cell function are caused by differential *expression* of the genes. Genes are expressed in a process where DNA is copied into RNA, which in turn guides the manufacturing of proteins. The natural control mechanisms of expression are largely unknown: some low-level building blocks have been determined, but uncovering regulatory networks of genes remains a big challenge.

DNA microarrays provide a technical method for measuring genetic expression. Single-stranded stretches of DNA are spotted onto a glass slide; each spot contains some DNA from a single gene, and is usually repeated at a few different locations on the slide. In the actual experiment, messenger RNA (mRNA) is obtained from the two tissue samples under consideration, complementary DNA (cDNA) is produced by a biochemical reverse transcription mechanism, and each of the two samples is labeled with a different fluorescent dye. These labeled cDNA samples are then allowed to hybridize on the slide, and the amounts of the two fluorescent dyes are measured with a confocal laser scanner. If a gene has differential expression in the two tissues, the original samples contain different amounts of mRNA, and thus different amounts of labeled cDNA ends up at the spot corresponding to the gene. This will cause a difference in the intensities measured by the scanner.

DNA microarrays can measure the expression levels of thousands of genes in parallel, whereas older methods require a separate experiment for each gene. They thus promise to help in the task of reverse engineering genetic regulation networks.

In medical genetics, the two compared tissues typically represent healthy and sick patients, for example normal lung cells and cancer cells. Also, e.g., the effects of a treatment can be monitored by using untreated tissue as reference and samples taken at different time points after the treatment; this yields time series data.

Experiments on Yeast Data

We examined several publically available datasets on expression in baker's yeast (*Saccharomyces cerevisiae*). 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 coregulated; but also genes far away from each other show a small but significant correlation. This could be explained by a hitherto unknown regulation mechanism or possibly by some chromosomally systematic bias in the microarray experiments. [1]

Human Data: Cancer and Asthma

We have collaborated with two research groups: one in the Centre of Biotechnology in Turku, led by Riitta Lahesmaa, and another in the Department of Medical Genetics at the University of Helsinki, led by Sakari Knuutila. The collaboration with Centre of

Biotechnology researchers has centered on validation of the Finnish microarray design intended for research into the genetic causes of asthma. We have given feedback on the quality of data obtained by researchers. This feedback has resulted in parts of some experiments being repeated on another scanner.

Together with the group of Sakari Knuutila, we have analyzed genetic expression data from patients suffering from mantle cell lymphoma [4], childhood leukemia [2], lung adenocarcinoma [3], and lung small cell carcinoma. The methods we have used have this far consisted of fairly standard statistics combined with computer-intensive permutation tests and ROC curve analyses. After these basic studies, we aim to continue with more ambitious clustering and other data analysis methods.

Gene Mapping

One of the large challenges in current medical genetics is locating genes predisposing to common multifactorial diseases. While there have been many successes in finding the genes responsible for various rare diseases, finding the genetic causes of common diseases such as asthma and diabetes poses still large problems.

Work done in colllaboration with the Finnish Genome Center and the Department of Computer Science of the University of Helsinki has developed methods for analyzing multifactorial diseases, for helping in study design, and in finding candidate loci for genes in multifactorial diseases.

One of the key problems in finding the genes predisposing to common traits is that the nature of genetic models is not well understood. A long-term goal in the group is to develop methods for deriving the potential genetic models from epidemiological data; the tools in this study are mainly Markov Chain Monte Carlo techniques. One recent result related to this study is a proof that any (single or multilocus) genetic model the sibling risk is at least as high as the offspring risk [5]. While the result is simple in the case of a single locus, the proof is highly nontrivial for the general case.

The design of empirical studies for finding genes contributing to certain traits is highly difficult. The population simulator developed in the same collaboration makes it possible to estimate whether a certain study design is appropriate for certain traits, and to estimate the significance of findings of studies; the tool has been used in a variety of studies [6, 7]. The haplotype pattern mining (HPM) method uses pattern discovery techniques to find the probable loci of predisposing genes. The performance of the method is extremely good [8], and it can also be generalized in a variety of ways [9].

References

[1] Heikki Mannila, Anne Patrikainen, Jouni K. Seppänen, and Juha Kere. Long-range control of expression in yeast. Accepted for publication in *Bioinformatics*.

[2] Tarja Niini, Jaakko Hollmén, Marcelo L Larramendy, Yan Aalto, Balint Nagy, Kim Vettenranta, and Sakari Knuutila. Gene Expression Profiling in Childhood Acute Lymphoblastic Leukaemia by cDNA array. Poster presented in *Beatson International Cancer Conference*, Genomic Regulation and Cancer, July 2001, Glasgow, Scotland.

[3] Harriet Wikman, Eeva Kettunen, Jouni K. Seppänen, Jaakko Hollmén, Antti Karjalainen, Sisko Anttila, and Sakari Knuutila. Gene Expression Profiling of Adenocarcinoma of the Lung. *Abstracts of the Asian-Pacific Conference on Tumor Biology*, September 2001, Peking, China.

[4] Ying Zhu, Jaakko Hollmén, Riikka Oinonen, Kaarle Franssila, Yan Aalto, Erkki Elonen, Heikki Mannila, Juha Kere, and Sakari Knuutila. Proffered paper presented in European Cancer Conference ECCO 11, October 2001, Lisbon. *European Journal of Cancer*, Vol. 37 Supplement 6, p. S41.

[5] M. Koivisto and H. Mannila. Offspring risk and sibling risk for multilocus traits. *Human Heredity* 51 (2001), 209–216.

[6] P. Kauppi, T. Laitinen, V. Ollikainen, H. Mannila, L.A. Laitinen, and J. Kere. The ILR9 region contribution in asthma is supported by genetic association in an isolated population. *European Journal of Human Genetics* 8, 788–792 (2000).

[7] T. Laitinen, V. Ollikainen, C. Lazaro, P. Kauppi, R de Cid, J.M. Anto, X. Estivill, H. Lokki, H. Mannila, L.A. Laitinen and J. Kere. Association study of the chromosomal region containing the FCER2 gene suggests it has a regulatory role in atopic disorders. *American Journal on Respiratory and Critical Care Medicine* 161, 700–706, 2000.

[8] H. Toivonen, P. Onkamo, K. Vasko, V. Ollikainen, P. Sevon, H. Mannila, M. Herr, and J. Kere. Data mining applied to linkage disequilibrium mapping. *American Journal of Human Genetics* 67(1): 133–145, July 2000.

[9] P. Sevon, V. Ollikainen, P. Onkamo, H. Toivonen, H. Mannila and J. Kere. Mining associations between genetic markers, phenotypes and covariates. *Genetic Epidemiology*, 21(Suppl 1): S588–S593, 2001.

Publications of the From Data to Knowledge Research Unit

Publications are in alphabetical order by the first author.

2001

1. Cadez, I., Smyth, P. & Mannila, H. Probabilistic Modeling of Transaction Data with Applications to Profiling, Visualization, and Prediction. *Proc. of 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2001)*, San Francisco, CA, USA, August 26-29, 2001. pp. 37-46.

2. Hand, D., Mannila, H. & Smyth, P. *Principles of Data Mining*. Boston 2001, MIT Press.

3. Himberg, J., Korpiaho, K., Mannila, H., Tikanmäki, J. & Toivonen, H. Time-Series Segmentation for Context Recognition in Mobile Devices. *Proc. of IEEE International Conference on Data Mining*, San Jose, CA, USA, Nov. 29 - Dec. 2, 2001. pp. 203-207.

4. Koivisto, M. & Mannila, H. Offspring Risk and Sibling Risk for Multilocus Traits. *Human Heredity*, 2001. Vol. 51, pp. 209-216.

5. Mannila, H. & Rusakov, D. Decomposition of Event Sequences into Independent Components. *First SIAM International Conference on Data Mining*, Chicago, IL, USA, April 5-7, 2001. http://www.siam.org/meetings/sdm01/pdf/sdm01_02.pdf

6. Mannila, H. & Salmenkivi, M. Finding Simple Intensity Descriptions from Event Sequence Data. *Proc. of 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2001)*, August 26-29, 2001, San Francisco, CA, USA. pp. 341-346.

7. Mannila, H. & Seppänen, J. K. Finding Similar Situations in Sequences of Events via Random Projections. *First SIAM International Conference on Data Mining*, Chicago, IL, USA, April 5-7, 2001.
http://www.siam.org/meetings/sdm01/pdf/sdm01_03.pdf

8. Toivonen, H.T.T., Mannila, H., Korhola, A. & Olander, H. Applying Bayesian Statistics to Organism-based Environmental Reconstruction. *Ecological Applications*, 2001. Vol. 11, No. 2, pp. 618-630.