Neural Information Processing

ICONIP2001 PROCEEDINGS

8th International Conference on Neural Information Processing

November 14–18, 2001
Shanghai, China

Volume 1

Edited by Liming Zhang and Fanji Gu

Fudan University Press

8th International Conference on Neural Information Processing

ICONIP2001

Copyright information

Fudan University Press
579 Guoquan Rd. Shanghai, 200433, China

(2001 Shanghai International Industrial FAIR Science Technology Forum)

Greetings From Steering Committee

Dear colleagues:

On behalf of the steering committee of ICONIP2001-Shanghai, I would like to extend to all the participants the sincerest greetings and warmest welcome.

Thanks to Prof. Amari and Fukushima from Japan, Prof. Wu from China, Prof. Lee from Korea and many other friends from AP region, ICONIP, earliest initiated during IJCNN1992-Beijing last century, has been a stable and strong platform for exchanging ideas, sharing experiences and exploring solutions among the scientists, professors and engineers in neural networks and the related fields within APNNA family.

The major goal of science and technology, to my understanding, is to find the good approaches to strengthening the human abilities among which the intelligence is the most demanding. Many evidences show that neural networks and the mechanisms of neural information processing would be the promising ones. On the other hand, however, the ever-made progress, up to the present stage, seems still very far from what ones expected. Hence there lies a long long way for us to go.

It has been getting more and more clear during the last decades that interdisciplinary cooperation would be much more effective than any other single disciplinary effort, even as neural networks, in the exploration of human intelligence as it may be the object with the highest complexity. As the result, it is a strong suggestion that a multi-disciplinary cooperation be specially encouraged among neural networks, artificial intelligence, biology, neurology, cognitive science, information science, knowledge theory, cybernetics, system theory, computer science, signal processing, and the like.

It is happy to see from the Proceedings of ICONIP2001 that this conference, like the previous ones, has implemented such multi-disciplinary cooperation. But it seemed not sufficient yet. It is our hope that more colleagues from various related fields and various regions in the world get together to make greater progress and more significant contributions to the human kinds.

Wish ICONIP2001-Shanghai a successful conference, wish the friendship within APNNA family everlasting, and, above all, wish all participants have an enjoyable and pleasant stay in Shanghai, China.

Professor Tian-De Shou
Professor Yi-Xin Zhong
Steering Committee Chairs

Message From The Program Committee

Dear Colleagues:

As Program Committee Chairmen of ICONIP2001, we are deeply honored to welcome you, our colleagues from all over the world, to the eighth International Conference on Neural Information Processing in Shanghai, China.

In the past few years, neural networks and biological motivated system have received a great deal of attention and are being touted one of the greatest computational tools. Many excitement results are due to the ability of neural networks to imitate the brain. Especially, the apparent ability to solve complex, noisy, nonlinear information processing problems that are difficult by other classical methods. It is true that neural information processing and brain science are closely integrated. The research areas on understanding brain, protecting brain and creating brain have pushed many scientists and experts in a variety of fields gathering on this annually meeting to share their ideas and new developments since 1994. It is also a great honor for China to hold the conference again in the famous and beautiful city, Shanghai.

We received over 300 submissions from more 31 countries and areas. With the help of renowned reviewers, the committee has selected 295 for oral and poster presentations. These papers with high scientific qualities, have covered brain model and cognitive science, learning algorithm, evolution and fuzzy system, neural network architecture, applications on image and signal processing, data mining, control, knowledge and rule extraction and others. Several leading scientists have been invited to give plenary presentation and forum talk on conference and sessions. Some of our colleagues help us to organize high-level mini-workshops and special sessions and some of our colleagues are also invited to chair the sessions. We deeply appreciate them for their kind support and help. We also give our thanks to the paper reviewers. We believed that each reviewer did his or her best to make objective decision though the review and decision process is never perfect. We apologize for any resulting disappointment.

Dear colleagues, on behalf of program committee, we would welcome you, the participants, to the ICONIP2001 and sincerely hope that conference successful!

Professor Aik Guo
Professor Mimory Tsukada
Professor Liming Zhang
Program Committee chairs

Conference Committees

Honorary Chairs

Shun-Ichi Amari, RIKEN BSI, Japan
Chao-Hao Gu, Fudan Univ., China

Conference Chairs

Kunihiko Fukushima, Tokyo Univ. of Technology, Japan
You-Shou Wu, Tsinghua Univ., China

Steering Committee Chairs

Tian-De Shou, Fudan Univ., China
Yi-Xing Zhong, China Post&Telecom. Univ., China

International Advisory Committee Chairs

Zhen-Ya He, Southeast Univ., China
Nikola K. Kasabov, Univ. of Otago, New Zealand

Members

Sung-Yang Bang, POSTECH, Korea
Tian-Ping Chen, Fudan Univ., China
Hui-Sheng Chi, Beijing Univ., China
Russell C. Eberhart, IUPUI, USA
T. Fukuda, Nagoya Univ., Japan
Tom Gedeon, The Univ. of New South Wales, Australia
San-Yan. Kung ,Princeton Univ. USA
Soo-Young Lee, KAIST, Korea
R. Marks, Univ. of Washington, USA
Erkki Oja, HUT, Finland
Tamas Roska Academy of Sciences, Hungarian
Harold Szu, George Washington Univ., USA
Shiro Usui, Toyohashi Univ. of Tech., Japan
Li-Po Wang, NTU, Singapore
Yun-Jiu Wang, Academia Sinica, China
P. Werbos, NSF, USA
Lei Xu, Chinese Univ. of Hong Kong, China
Jacek Zurada, Univ. of Louisville, USA

Organizing Committee Chairs

Fan-Ji Gu, Fudan Univ., China
Meng-Qi Zhou, CNNC, China

Finance Chair

Yong-De Shi, Fudan Univ., China

Forty Years of Perceptrons

Shun-ichi Amari

RIKEN Brain Science Institute
Wako-shi, Hirosawa 2-1, Saitama 351-0198, Japan
amari@brain.riken.go.jp

Abstract

More than forty years have passed since a learning machine called the perceptron was proposed by Rosenblatt. It has played a major role for these forty years, not only in the area of computational neuroscience in elucidating the function of cerebellum, but also in theoretical and practical studies of learning machines. The present talk reviews these developments from my personal point of view.

1 Perceptron and Its Convergence Theorem

Perceptron is a model of neural networks having learning ability. It was proposed by Rosenblatt in the late fifties, where many types of perceptron were studied Rosenblatt [31]. They included multilayer perceptrons having laterally connected and feedback connected ones, and even a possibility of error signals backpropagating was stated.

The simple perceptron was studied deeply. It consists of three layers, the input layer, the association layer and the output layer. The connections from the input to the association layers are randomly assigned so that the perceptron can perform any tasks universally. The connections from the association to the output layers are modifiable, so that it can be trained to perform any specific task by learning from examples.

The perceptron convergence theorem was proved by [14] which guarantees that any linearly separable function can be realized by a simple perceptron by a finite number of training from examples. The proof was very complicated, and simpler proofs were given by Novikov and by Minsky and Papert [27]. The simple perceptron had been a central topic of research in the early sixties in the community of learning, threshold logic and pattern recognition.

Unfortunately, other types of perceptrons were not well studied, because they were so difficult to analyze theoretically, and computers at that time were so poor to do large-scale simulations. Minsky and Papert [27] studied the capability of the simple perceptron and condemned that its performance is severely limited from the point of view of computational complexity. Their theory is fundamental and interesting, although the simple perceptron is too simple as a model of neural parallel processing. It was said to be unfortunate that the theory had influence on the decay of studies of perceptrons in the seventies. However, this story is only a myth, and it was not Minsky-Papert paper (book) that caused the decay of neural modelling. Computers became much more powerful in that period, and a more practical approach to pattern recognition as well as the AI type research had become much more fashionable at that time, attracting many researchers. However, there were important developments outside America and Europe in the late sixties and early seventies, which has not yet been well recognized. Before describing that story, a biological impact of perceptron will be stated.

2 Perceptron Theory of Cerebellum

The cerebellum has a well-structured multilayer architecture, which was studied by experimentalists in the sixties. They asked theoreticians what was implied by such a structure theoretically in information processing. It was Marr [26] who answered the question, by proposing the perceptron theory of cerebellum. The mossy fibers were identified as the inputs to the cerebellum, and the granular cells receive inputs from mossy fibers by

random connections. The granular cells form the association layer, sending signals to the output Purkinje cells through parallel fibers. These connections are modifiable, where the teacher signals are provided from the climbing fibers. A similar idea was also proposed by Albus [2] independently.

The theory gave an impact to experimentalists, but it was impossible to prove the idea by experiments at that time. So most experimentalists were not in favor of the theory, neglecting its implications. Ito [22] was one of a few who took the idea seriously, and designed a thoughtful experimental paradigm to prove that neural learning in cerebellum takes place in the synapses of the parallel fibers to the Purkinje cells, where the teacher signals are supplied from the climbing fibers. It took ten years of tireless efforts, before he proved the long term depression taking place in the connections of the Purkinje cells. Long term potentiation was found in hippocampus, which was the first evidence of plasticitic changes in the brain. Ito's experiment showed the long term depression. This not only proved another type of plastic change in the brain, but more importantly provided evidence of neural plasticity connected with the change in the behavior of animals. That is, this was the first finding that behavioral changes in mammals are caused by neural plastic changes in synapses.

This finding shows the importance and fruitfulness of cooperation between theoretical and experimental research. However, there have been a number of experimentalists who were strongly against Ito's findings. They proposed alternative ideas of learning against Ito. Whenever their ideas were disproved, they proposed another one, and such process has been repeated, in spite that majority has approved Ito's idea.

3 Stochastic Descent Learning

In the period of the sixties, Widrow [39] proposed another learning machine called the "adaline". This is an adaptive linear neuron, which performs learning by modifying the connection weights in the direction of the gradient of the squared error. This is the origin of the stochastic descent learning method. The adaline is linear, so that it cannot be generalized to multilayer nonlinear machines. The perceptron learning by Rosenblatt was proposed from a completely different point of view, because it uses binary neurons to perform logical calculus, so that the stochastic descent method cannot be applied to multilayer perceptrons in the original form.

In the late sixties, there were new developments in learning methods. One is in Russian school. Aizerman et al [1] proposed the potential function method of learning, which is a linear machine using the stochastic descent. Input signals are transformed nonlinearly so that its capability is universal. This is an origin of the kernel support vector machine (Scholkopf et al [34]). Tsypkin [36] applied the method of stochastic approximation for learning. Vapnik also proposed the idea of VC dimensions in relation to the uniform convergence of learning machines.

Amari [3] studied the dynamic behavior of learning, and analyzed the relation between the learning speed and accuracy. To this end, he used an analog sigmoidal activation function, and formulated the learning rule in the framework of stochastic gradient descent. He then applied it to learning of multilayer perceptrons, which was later named the delta rule by Rumerhalt et al [32]. Amari did not find the interesting interpretation that the error signal backpropagates in multilayer perceptrons, from which the name of backpropagation came (Rumerhalt et al [32]). However, his framework was more general but the algorithm was essentially the same as the rediscovered one. Moreover, the dynamics of online learning, the speed and accuracy, was analyzed. The same results were rediscovered later by Heskes and Kappen [20] by statistical physical method, and became a source of further research on online learning (Saad [33]). The adaptive control of the learning rate was also suggested in the old paper.

It was strange that all of these works have been completely forgotten and ignored in the later developments of neural networks in the eighties. There was a strong group of machine learning in Moscow as mentioned, and they were surprised to find similar research in Japan, because these types of research were scarce in the US and in Europe at that

time (personal communication by Vapnik). When Tsypkin [36] wrote a book on pattern recognition and learning in the seventies, he devoted one chapter to an introduction of the theory of Amari [3]. See also [5].

4 New Developments of Perceptron

The (re)discovery of backpropagation (Rumelhart et al [32]) was welcomed enthusiastically by the neural networks community. Computers were sufficiently matured in this time so that large-scale simulations were possible. The model "Netalk" by Sejnowski and Rosenberg [35] was one of the earliest results which demonstrated the power and usefulness of the multilayer perceptrons as practical engineering tools. Although the multilayer perceptron includes difficulties in slow convergence and existence of local minima, it has been fully developed to be a standard engineering tool for a learning pattern recognizer and has been applied in various fields of science and engineering.

Theoretical studies have further been developed rapidly. It was proved that the three-layer perceptron is a universal approximator of nonlinear functions in the sense that any function can be approximated by it, provided it includes a sufficiently large number of hidden units (Funahashi [16], Hornik et al [21]). More surprisingly, it was proved by Jones [23] and by Barron [13] that the multilayer perceptron is free from the "curse of dimensionality". In conventional methods of function approximation, when the dimensions of the input signals are large, the number of parameters to be adjusted for approximating a nonlinear function of inputs signals increases exponentially. However, when a function to be approximated is smooth, the number does not increase exponentially in the case of perceptrons, implying that the perceptron is free from the curse of dimensionality. This surprising result was proved by the information-theoretic method.

There have been developed statistical (Amari and Murata [9]) as well as statistical physical theories of learning in such machines (Levin et al [24]; Saad [33]). The relation between training error and generalization error was elucidated by these methods. Bayesian theories and the regularization theory were also developed (MacKay [25]: Poggio and Girosi [29]). Vapnik used the theory of uniform convergence to develop a statistical learning theory, which elucidated the relation between training error and generalization error from a more fundamental point of view. This resulted in the proposal of the support vector machines and kernel support vector machines, which have a strong power in pattern recognition and nonlinear regression. Bagging and boosting methods came from a different statistical viewpoint, which overcomes the problem of local minima.

Machine learning has become a large interdisciplinary area of research where neural networks, pattern recognition [34], artificial intelligence, statistics, statistical physics and many other disciplines are merged.

5 Information Geometry of Multilayer Perceptrons

Information geometry (Amari and Nagaoka [10]) originated from the study of intrinsic properties existing in the manifold of probability distributions. It has successfully been applied not only to the theory of statistical inference but also in many other fields such as information theory [18], control systems theory [4] and neural networks [8], signal processing such as independent component analysis [7]. In particular, information geometry of multilayer perceptrons has been developed, which opened a way to new interesting aspects of hierarchical systems including multilayer perceptrons.

The set of multilayer perceptrons forms a geometrical manifold where the set of modifiable parameters (connection weights and thresholds) play the role of a coordinate system. When learning takes place, it is represented by a trajectory in the manifold. By taking the noise into account, the behavior of a perceptron is described by a conditional probability distribution of the output conditioned on the input. Hence, the parameter space of perceptrons, which we call the neuromanifold, consists of a set of conditional probability distributions, which is a topic elucidated by information geometry.

The neuromanifold is a Riemannian manifold having the Fisher information matrix as its metric structure. When such a structure is taken into account, the conventional gradient should be replaced by the Riemannian gradient or natural gradient (Amari [6]). If the neuromanifold is not so strongly curved, there are only small difference between the conventional and the natural gradients. However, because of the symmetry existing in such hierarchical systems, it is strongly curved. Moreover, it includes singularities where the Fisher information degenerates.

It was shown that the plateau phenomena by which perceptron learning becomes very slow are given rise to by such singular geometrical structures. Amari proposed natural gradient learning to overcome this difficulty. However, the calculation of the inverse of the Fisher information matrix seemed rather difficult. This difficulty was overcome by the adaptive natural gradient method (Amari et al [12], Park et al [28]). Simulations show its extremely quick convergence ability in learning. Statistical physical analysis also confirmed this (Rattray et al [30]).

The success of the natural gradient learning method posed another interesting questions: How a learning trajectory is affected by such singularities? (Amari and Ozeki [11]) More generally, we need to elucidate learning and statistical inference in the presence of singularities in neuromanifolds or statistical models in general. Watanabe [38] used modern algebraic geometry to elucidate this problem. This is now a very important new topic of research, developed by a number of Japanese researchers [15], [17], [19].

6 Conclusions

The present paper reviewed forty years of developments of perceptrons. It is interesting to see that the concept of perceptrons generated a lot of new ideas, including biological, theoretical and practical applications throughout forty years. It is still a source of new ideas.

References

[1] Aizerman, M.A., Braverman, E.M. & Rozonoer, L.I. *The probability problem of pattern recognition learning and the method of potential functions*, Automation and Remote Control, Vol. 25, pp. 243-247, 1964.

[2] Albus, J.S. *A theory of cerebellar function*, Math. Biosciences, Vol. 10, pp. 25-61, 1971.

[3] Amari, S. *Theory of Adaptive Pattern Classifiers*, IEEE Trans., EC-16, No. 3, pp. 299-307, 1967.

[4] Amari, S. *Differential geometry of a parametric family of invertible linear systems - Riemannian metric, dual affine connections and divergence*, Mathematical Systems Theory, Vol. 20, pp. 53-82, 1987.

[5] Amari, S. *Backpropagation and Stochastic Gradient Descent Method, Neurocomputing*, Vol. 5, No. 4-5, pp. 185-196, Elsevier, 1993.

[6] Amari, S. *Natural Gradient Works Efficiently in Learning*, Neural Computation, 10, pp. 251-276, 1998.

[7] Amari, S. *Superefficiency in Blind Source Separation*, IEEE Transactions on Signal Processing, Vol. 47, No. 4, pp. 936-944, 1999.

[8] Amari, S., Kurata, K. & Nagaoka, H. *Information Geometry of Boltzmann Machines*, IEEE Trans. on Neural Networks, Vol. 3, No. 2, pp. 260-271, 1992.

[9] Amari, S. & Murata, N. *Statistical theory of learning curves under entropic loss criterion*, Neural Computation, Vol. 5, pp. 140-153, 1992.

[10] Amari, S. & Nagaoka, H. *Methods of Information Geometry, Translations of Mathematical Monographs*, Vol. 191, Oxford University Press, 2000.

[11] Amari, S. & Ozeki, T. *Differential and Algebraic Geometry of Multilayer Perceptrons*, IEICE Trans. Fundamentals, Vol. E84-A, No. 1, pp. 31-38, 2001.

[12] Amari, S., Park, H-Y. & Fukumizu, K. *Adaptive Method of Realizing Natural Gradient Learning for Multilayer Perceptrons*, Neural Computation, Vol. 12, pp. 1399-1409, 2000.

[13] Barron, A.R. *Universal approximation bounds for superpositions of a sigmoidal function*, IEEE Transactions on Information Theory, Vol. 39, pp. 930-945, 1993.

[14] Block, H.D. *The perceptron, a model for brain functioning I*, Rev. of Modern Physics, Vol. 34,

pp. 123-135, 1962.

[15] Fukumizu, K. *Generalization error of linear neural networks in unidentifiable cases*, Lecture Notes in Computer Sciences, 1720, eds. O. Watanabe and T. Yokomori, pp. 51-62, Springer, 1999.

[16] Funahasi, K. *On the approximate realization of continuous mappings by neural networks*, Neural Networks, Vol. 2, pp. 183-191, 1989.

[17] Hagiwara, K., Toda, N. & Usui, S. *Nonuniqueness of connecting weights and AIC in multi-layered neural networks*, IEICE Trans., Vol. J76-D-II, pp. 2058-2065, 1993.

[18] Han, T.S. &Amari, S. *Statistical Inference Under Multiterminal Data Compression*, IEEE Transactions on Information Theory, Vol. 44, No. 6, pp. 2300-2324, 1998.

[19] Hayasaka, T., Hagiwara, Ktoda, N. & Usui, S. *On the estimation and prediction errors of function representation with orthonormal discrete variable basis in regression model*, The Brain and Neural Networks, Vol. 4, pp8-26, 1997.

[20] Heskes, T.M. & Kappen, B. *Learning processes in neural networks*, Physical Review, A, Vol. 44, pp. 2718-2726, 1991.

[21] Hornik, K., Stinchcombe, M., & White, H. *Multilayer feedforward netwroks are universal approximators*, Neural Networks, Vol. 2, pp. 359-366, 1989.

[22] Ito, M. *The Cerebellum and Neural Control*, Raven Press, 1984.

[23] Jones, L.K. *Constructive approximations for neural networks by sigmoidal functions*, Proceedings of IEEE, Vol. 78, pp. 1586-1589, 1990.

[24] Levin, E., Tishby, N. & Solla, S.A. *A statistical approach to learning and generalization in layered neural networks*, Proceedings of IEEE, Vol. 78, pp. 1568-1574, 1990.

[25] MacKay, D. *A practical Bayesian framework for backpropagation networks*, Neural Computation, Vol. 4, pp. 448-472, 1992.

[26] Marr, D. *A theory of cerebellar cortex*, J. Physiol., Vol. 202, pp. 437-470, 1969.

[27] Minsky, M. & Papert, S. *Perceptron-An Essay in Computational Geometry*, MIT Press, 1969.

[28] Park, H., Amari, S. & Fukumizu, K. *Adaptive natural gradient learning algorithms for various stochastic models*, Neural Networks, Vol. 13, pp. 755-764, 2000.

[29] Poggio, T. & Girosi, F. *Networks for approximation and learning*, Proceedings of IEEE, Vol. 78, pp. 1481-1497, 1990.

[30] Rattray, M, Saad, D. & Amari, S. *Natural Gradient Descent for On-Line Learning*, Physical Review Letters, Vol. 81, No. 24, pp. 5461-5464, 1998.

[31] Rosenblatt, F. *Principles of Neurodynamics*, Spartan, 1961.

[32] Rumelhart, D.E. Hinton G.E. & Williams, R.J. *Learning internal representation by error backpropagation,* in Parallel Distributed Processing (D.E.Rumelhart and J.L.McClleland eds.), MIT Press, pp. 318-362, 1986.

[33] Saad, D. *On-Line Learning in Neural Networks*, Cambridge University Press, 1998.

[34] Scholkopf, B., Burges, C.J.C. & Smola, A.J. *Advances in Kernel Methods: Support Vector Machines*. MIT Press, 1998.

[35] Sejnowski T.J. & Rosenberg, C.R. *Parallel networks that learn to pronounce English text*, Complex Systems, Vol. 1, pp. 145-168, 1987.

[36] Tsypkin, Y. *Foundations of the Theory of Learning Systems*, Academic Press, 1973.

[37] Vapnik, V.N. *Statistical Learning Theory*, John Wiley, 1998.

[38] Watanabe, S. *Algebraic analysis for non-identifiable learning machines*, Neural Computation, Vol. 13, pp. 899-933, 2001.

[39] Widraw, B. *A Statistical Theory of Adaptation*, Pergamon Press, 1963.

Modulation of Visual Information Processing Through Various Retinal Signal Channels

Xiong-Li Yang

Institute of Neurobiology, Fudan University
220 Han-Dan Road, Shanghai 200433, China

Abstract

The vertebrate retina, an approachable part of the brain, is composed of just six basic nerve cells, which are organized into several clearly distinct cellular layers, thus providing a good model for exploring the mechanisms underlying brain function. It has long been recognized that there are parallel streams of preprocessed information passing from the retina to the brain for higher perceptual processing, via a variety of channels for visual contrast, visual adaptation, color, special frequency. In this lecture, general principles of signal transfer along the channels in the retina will be summarized and the data will be presented to demonstrate how the transmission through the "red-cone" channel, the chromatic channel and the "on-off" channel is modulated by darkness, illumination and a variety of chemicals, including neurotransmitters and neuromodulators. Possible mechanisms underlying the modulation and hints to neuroscientists, working in the field of neural network and nervous system modeling which these studies may provide are discussed as well.

FINDING STRUCTURE IN SIGNALS, IMAGES, AND DATA

Erkki Oja

Helsinki University of Technology
Neural Networks Research Centre
P.O. Box 5400, FIN-02015 HUT, Finland
erkki.oja@hut.fi

ABSTRACT

The talk will be a tutorial survey, concentrating on the main principles and categories of unsupervised neural learning in the problem of data mining for signals, images, and data. In neural computation, there are two classical categories for unsupervised learning methods and models: first, extensions of Principal Component Analysis and Factor Analysis, and second, learning vector coding or clustering methods that are based on competitive learning. The talk concentrates on two of these extensions: for the first category, the novel technique of Independent Component Analysis, and for the second category, the Kohonen Self-Organizing Map. The more recent trend in unsupervised learning is to consider this problem in the framework of probabilistic generative models. If it is possible to build and estimate a model that explains the data in terms of some latent variables, key insights may be obtained into the true nature and structure of the data. This approach is also briefly reviewed. After a brief introduction to the underlying theoretical foundations of these ideas, unsupervised neural learning will be illustrated by several applications in data mining ranging from document and pictorial databases to blind signal separation.

Keywords unsupervised learning, data mining, independent component analysis, self-organizing map

Acknowledgement

This work was supported by the Finnish Centre of Excellence Programme (2000-2005) of the Academy of Finland, project New information processing principles, 44886.

1. INTRODUCTION

Progress in computer and information sciences was for a long time restricted by the state-of-the-art of computer hardware and data networks. In recent years a new situation has been encountered: the worldwide proliferation of powerful computing services has caused an uncontrolled flood of information in the Internet and other media. It therefore becomes increasingly important to develop fundamentally new information processing principles for making relevant knowledge accessible to the user and to present it in a comprehensible form. This means, for example, completely new explorative data analysis and data mining methods, combined with advanced graphics facilities.

Along with the explosive increase in available digital data, the computing power of modern hardware has been dramatically increased as well. With the increasing computing power, it has become possible to digitally process and classify huge masses of natural data, such as statistical information, images, speech, text, as well as other kinds of signals and measurements coming from very different sources. Such tasks occur in industry, remote sensing, medicine, finance, and natural sciences, to mention only a few main fields. For financial, medical, administrative, and other databases, one needs efficient tools for visualization, prediction, clustering, and profiling. In industrial problems, it is essential to build empirical data based models of complex systems in order to be able to predict, monitor, diagnose faults, and control the systems.

One of the central tools in data mining is unsupervised learning. This means a completely data driven approach in which the pertinent structure, in the form of patterns, clusters, or models, is automatically found from the data using advanced statistical and computational techniques. Some insight into the unsupervised learning problem can be inferred from cognitive science. It is obvious that many effective computing principles that we do not yet know in detail exist in the biological brain. For example, many hierarchical computing structures of the brain have still remained a mystery. On the other hand, the mathematical operations and expressions that we use for the description of known neural operations can be computed digitally with much higher accuracy and stability than what is possible by the analog computing principles of the biological networks. Therefore, trying to combine the best of these two worlds is a strong motivation, emerging in the research field of neural computation. This can be seen as being situated at the intersection

of machine learning, computation, and advanced statistics.

The Section 2 of this paper reviews the three main approaches to unsupervised machine learning in neural networks. Then, Section 3 illustrates these approaches by some well-known concrete mathematical models. Section 4 mentions some applications that will be covered in detail in the talk.

This paper is based on the more extensive review (Oja, 2001).

2. WHAT IS UNSUPERVISED LEARNING

Unsupervised learning is a deep concept that can be approached from very different perspectives, from psychology and cognitive science to engineering. It is often called "learning without a teacher". This implies that a learning human, animal, or artificial system observes its surroundings and, based on these observations, adapts its behavior without being told how to associate given observations to given desired responses (supervised learning) or without even given any hints about the goodness of a given response (reinforcement learning). Usually, the result of unsupervised learning is a new explanation or representation of the observation data, which will then lead to improved future responses or decisions (Hinton and Sejnowski, 1999). This is precisely the problem in data mining, too.

In machine learning and artificial intelligence, such a representation is a set of concepts and rules between these concepts, which give a symbolic explanation for the data. In advanced statistics, the representation may be a clustering of the data, a discrete map, or a continuous lower-dimensional manifold in the vector space of observations, which explains their structure and may reveal their underlying causes.

Unsupervised learning seems to be the basic mechanism for sensory adaptation e.g. in the visual pathway (Barlow, 1989). If we accept the hypothesis that biological learning is based on synaptic modification, a big problem is how supervised learning rules like back-propagation could be implemented locally on the synaptic level. The biological substrate seems to be much more compatible with the unsupervised mode of learning. For more biologically oriented neural approaches, see (Grossberg, 1988). On the engineering side, unsupervised learning is a highly powerful and promising approach to some practical data processing problems like data mining and knowledge discovery from very large databases, or new modes of human-computer interactions in which the software adapts to the requirements and habits of the human user by observing her behaviour.

3. EXAMPLES OF UNSUPERVISED LEARNING IN NEURAL COMPUTATION

In neural computation, there have been two classical categories for unsupervised learning methods and models: first, extensions of Principal Component Analysis and Factor Analysis, and second, learning vector coding or clustering methods that are based on competitive learning (Haykin, 1999). The more recent trend in unsupervised machine learning is to consider this problem in the framework of probabilistic generative models (Hinton and Sejnowski, 1999). If it is possible to build and estimate a model that explains the data in terms of some latent variables, key insights may be obtained into the true nature and structure of the data. Operations like prediction and compression become easier and rigorously justifiable.

3.1. The Self-Organizing Map

The goal of unsupervised learning, finding a new compressed representation for the observations, can be interpreted as coding of the data. Thus learning vector coding methods that are based on competitive learning can be highly useful. A typical application is data mining or profiling from massive databases. It is of interest to find out what kind of typical clusters there are among the data records. In a customer profiling application, finding the clusters from a large customer database means more sharply targeted marketing with less cost. In process modelling, finding the relevant clusters of the process state vector in real operation helps in diagnosis and control. A competitive learning neural network gives an efficient solution to this problem. The best-known competitive learning network is the Self-Organizing Map (SOM) introduced by Kohonen (see Kohonen, 2001).

In vector coding, the problem is to place a fixed number of vectors, called *codewords*, into the input space which is usually a high-dimensional vector space. The input data (observations) are given as a training set of numerical vectors $\mathbf{x}(1), ..., \mathbf{x}(T)$. For example, the inputs can be gray-scale windows from a digital image, measurements from a machine or an industrial process, financial data describing a company or a customer, or pieces of English text represented by word histograms. The dimension n of the data vectors is determined by the problem and can be very large. In the WEBSOM system for organizing collections of text documents (Kohonen *et al*, 2000), the dimensionality of the data in the largest applications is about $n = 50,000$ and the size of the training sample is about $T = 7,000,000$.

The goal of SOM learning is not only to find the most representative code vectors for the input training set in the sense of minimum distance, as is the case in the usual vector coding methods, but at the same time to form a topological mapping from the input space to the grid or lattice of neurons. This idea originally stems from the modelling

of the topographic maps on the sensory cortical areas of the brain. A related early work in neural modelling is (Malsburg, 1973).

For any data point $\mathbf{x}$ in the input space, one or several of the codewords are closest to it. Assume that $\mathbf{w}_i$ is the closest among all codewords:

$$||\mathbf{x} - \mathbf{w}_i|| = min||\mathbf{x} - \mathbf{w}_j||, j = 1, ..., k \quad (1)$$

The unit i having the weight vector $\mathbf{w}_i$ is then called the *best-matching unit* (BMU) for vector $\mathbf{x}$. Note that for fixed $\mathbf{x}$, Eq. (1) defines the index $i = i(\mathbf{x})$ of the BMU, and for fixed i, Eq. (1) defines the set of points $\mathbf{x}$ that are mapped to that index and thus all belong to the same cluster. By the above relation, the input vectors $\mathbf{x}$ are mapped to the discrete set of indeces i.

By a topological mapping the following property is meant: if a given point $\mathbf{x}$ is mapped to unit i, then all points in neighborhoods of $\mathbf{x}$ are mapped either to i itself or to one of the units in the neighborhood of i in the lattice. Because no topological maps between two spaces of different dimensions can exist in the strict mathematical sense, a two-dimensional neural layer can only follow locally two dimensions of the multidimensional input space. Usually the input space has a much higher dimension, but the data cloud $\mathbf{x}(1), ..., \mathbf{x}(T)$ used in training may be roughly concentrated on a lower-dimensional manifold that the map is able to follow at least approximately (Kohonen, 2001). The well-known Kohonen learning rule is able to tune the map so that weight vectors attain optimal positions. For recent advances on the SOM, see (Oja and Kaski, 1999).

3.2. PCA, ICA, and FA

The other class of unsupervised learning methods is motivated by standard statistical methods like Principal Component Analysis (PCA) or Factor Analysis (FA), which give a reduced subset of linear combinations of the original input variables. A classical approach are the on-line PCA learning rules introduced by the author (Oja, 1982). As an example, consider here Factor Analysis (see e.g. Harman, 1967).

In FA, a generative latent variable model is assumed for the observation vectors $\mathbf{x}$:

$$\mathbf{x} = \mathbf{A}\mathbf{y} + \mathbf{n}. \quad (2)$$

FA was originally developed in social sciences and psychology. In these disciplines, the researchers want to find relevant and meaningful factors that explain observed results. The interpretation in the model (2) is that the elements of $\mathbf{y}$ are the *unobservable* factors. The elements a_{ij} of the unknown matrix $\mathbf{A}$ are called *factor loadings*. The elements of the unknown additive term $\mathbf{n}$ are called specific factors. The elements of $\mathbf{y}$ (the factors) are uncorrelated, zero mean and gaussian, and their variances are absorbed into the matrix $\mathbf{A}$ so that we may assume

$$\mathrm{E}\{\mathbf{y}\mathbf{y}^T\} = \mathbf{I}. \quad (3)$$

The elements of vector $\mathbf{n}$ are zero mean, uncorrelated with each other and also with the factors y_i; denote $\mathbf{Q} = \mathrm{E}\{\mathbf{n}\mathbf{n}^T\}$. It is a diagonal matrix. We may write the covariance matrix of the observations from (2) as

$$\mathrm{E}\{\mathbf{x}\mathbf{x}^T\} = \mathbf{C}_\mathbf{x} = \mathbf{A}\mathbf{A}^T + \mathbf{Q}. \quad (4)$$

In practice, we have a good estimate of $\mathbf{C}_\mathbf{x}$ available, given by the sample covariance matrix. The main problem is then to solve the matrix $\mathbf{A}$ of factor loadings and the diagonal covariance matrix $\mathbf{Q}$ such that they will explain the observed covariances from (4). There is no closed-form analytic solution for $\mathbf{A}$ and $\mathbf{Q}$. Assuming $\mathbf{Q}$ is known or can be estimated, we can solve $\mathbf{A}$ from $\mathbf{A}\mathbf{A}^T = \mathbf{C}_\mathbf{x} - \mathbf{Q}$. This solution is not unique, however: any matrix $\mathbf{A}' = \mathbf{A}\mathbf{T}$ where $\mathbf{T}$ is an orthogonal matrix ($\mathbf{T}\mathbf{T}^T = \mathbf{I}$) will also be a solution. Then the factors will change to $\mathbf{y}' = \mathbf{T}^T\mathbf{y}$. For $\mathbf{A}'$ and $\mathbf{y}'$, the FA model (2) holds, and the elements of $\mathbf{y}'$ are still uncorrelated. The reason is that the property of uncorrelatedness is invariant to orthogonal transformations (rotations). Note that because the factors are uncorrelated and gaussian, they are also independent.

In *Independent Component Analysis* (ICA) (see e.g Amari, 1996; Bell and Sejnowski, 1995; Cardoso, 1998; Hyvärinen, Karhunen and Oja, 2001; Jutten, 1991), the same model (2) is assumed, but now the assumption on y_i is much stronger: we require that they are *statistically independent* and *non-gaussian*. Interestingly, then the ambiguity in Factor Analysis disappears and the solution, if we can find one, is (almost) unique.

In the simplest form of ICA, the additive noise $\mathbf{n}$ is not included and the standard notation for the independent components or *sources* is s_i; thus the ICA model for observation vectors $\mathbf{x}$ is

$$\mathbf{x} = \mathbf{A}\mathbf{s}. \quad (5)$$

It is again assumed that both $\mathbf{x}$ and $\mathbf{s}$ are zero mean. The observations x_i are now linear combinations or mixtures of the sources s_j. The matrix $\mathbf{A}$ is called in ICA the *mixing matrix*. In a typical application of ICA, a set of parallel time signals such as speech waveforms, electromagnetic measurements from the brain, or financial time series, are assumed to be linear combinations of underlying independent latent variables. The variables, which are now the independent components, are found by efficient ICA learning rules.

A recent survey on ICA is (Hyvärinen, Karhunen and Oja, 2001) that also contains an extensive list of citations to the original literature.

3.3. Nonlinear Generative Models

The concept of a generative model is very general and potentially powerful. In fact, as discussed by (Roweis and Ghahramani, 1999), a large number of central techniques like FA, PCA, ICA, mixtures of Gaussians, vector quantization, and also dynamical models like Kalman filters or Hidden Markov Models, can be presented in a unified framework of unsupervised learning under a single basic generative model. In the Bayesian Ying - Yang model (Xu 2000), likewise a generic framework of unsupervised learning is employed for the basic data models, both static and temporal.

We already saw examples of generative models in the case of Factor Analysis and Independent Component Analysis. Also Principal Component Analysis can be derived from a generative model in the technique called Probabilistic PCA (Tipping and Bishop, 1999). A problem with such linear models, however, is that they cannot represent well data that is not a linear mixture of some underlying gaussian or nongaussian variables. For data clouds that have an irregular or curved shape, these methods fail.

In the Generative Topographic Map (GTM) algorithm (Bishop *et al*, 1998), the observation vectors $\mathbf{x}$ are expressed in terms of a number of latent variables, which are defined on a similar lattice or grid as the neurons in the SOM. The mapping from the latent variables to the observations is *nonlinear*:

$$\mathbf{x} = \mathbf{f}(\mathbf{y}, \mathbf{M}) + \mathbf{n} \tag{6}$$

where $\mathbf{M}$ is an array of parameters of the nonlinear function $\mathbf{f}$, and $\mathbf{n}$ is additive noise. The form of the function $\mathbf{f}$ is assumed to be determined except for the unknown parameters. The model (6) is the generative latent variable model of the GTM method. It means that the observed data vectors $\mathbf{x}$ are basically concentrated on a lower dimensional nonlinear manifold in the data space, except for the additive noise. The vectors $\mathbf{w}_i = \mathbf{f}(\mathbf{y}_i, \mathbf{M})$ that are the images of the node points $\mathbf{y}_i$ are analogous to the weight vectors or codewords of the SOM. If $\mathbf{f}$ is smooth, a topographic ordering for the codewords is automatically guaranteed, if such an ordering is valid for the latent points $\mathbf{y}_i$. The GTM also has the advantage that it postulates a smooth manifold that naturally interpolates between the code vectors $\mathbf{w}_i$. The parameters can be learned using the EM algorithm.

When comparing the FA model (2) and the GTM model (6), certain similarities emerge: both have a number of latent variables, given by the vector $\mathbf{y}$, and additive gaussian noise $\mathbf{n}$. In FA, the mapping from $\mathbf{y}$ to the data $\mathbf{x}$ is linear, in GTM it is nonlinear. Another clear difference is that in FA, the factors are gaussian, while in GTM, the prior density $p(\mathbf{y})$ for the latent factors has a very special (atomic) form.

Another possibility for this density in the nonlinear case, too, would be the gaussian density, which would then be close to the original flavor of FA. If we assume that the prior for $\mathbf{y}$ is gaussian with unit (or diagonal) covariance, making the elements y_i independent, as in eq. (3), then the model (6) may be called *nonlinear factor analysis*. A further extension would be $p(\mathbf{y})$ that is *nongaussian but factorizable* so that the y_i are independent; then the model becomes *nonlinear independent component analysis*.

Recently, (Valpola, 2000) used an approximation for the nonlinear function $\mathbf{f}(\mathbf{y}, \mathbf{M})$ in the model, that was based on a Multilayer Perceptron (MLP) network with one hidden layer. It is well-known (see e.g. Haykin, 1998) that this function can approximate uniformly any continuous functions on compact input domains and it is therefore suitable for this task. Then the model becomes

$$\mathbf{x} = \mathbf{B}\phi(\mathbf{A}\mathbf{y} + \mathbf{a}) + \mathbf{b} + \mathbf{n} \tag{7}$$

where $\mathbf{A}, \mathbf{a}$ are the weight matrix and offset vector of the hidden layer, ϕ is the sigmoidal nonlinearity, typically a tanh or $\sinh^{-1}$ function, and $\mathbf{B}, \mathbf{b}$ are the weight matrix and offset vector of the linear output layer. It is understood that ϕ is applied to its argument vector element by element. In practice, there is a training sample $\mathbf{x}(1), ..., \mathbf{x}(T)$, and we wish to solve from the model the corresponding source or factor vectors $\mathbf{y}(1), ..., \mathbf{y}(T)$.

The problem now is that, contrary to the usual supervised learning situations, the inputs to the MLP are not known and therefore back-propagation type of learning rules cannot be used for finding the unknown parameters. The idea in (Valpola, 2000) is to use a purely Bayesian approach called *ensemble learning*. The cost function is the Kullback - Leibler divergence between the true posterior probability for the parameters, given the observations, and an approximation of that density. Several applications with real data have ben shown. The model has also been extended to a dynamical model, similar to an extended Kalman filter but with unknown parameters, and very promising results are obtained in case studies.

4. APPLICATIONS

The talk will be an introductory survey, concentrating on the main principles and categories of unsupervised learning. In the talk, the theoretical foundations of unsupervised machine learning will be shortly reviewed and the techniques will be illustrated by several applications in data mining: finding relevant documents in large document collections, content-based image retrieval, finding structure in biomedical measurements, and finding hidden nonlinear factors in time series. For more information and references, see the Web pages (NNRC, 2001).

References

Amari, S.- I., Cichocki, A. and Yang, H., "A new learning algorithm for blind source separation". In *Advances in Neural Information Processing Systems 8*, Cambridge: MIT Press, 1996, pp. 757 - 763.

Barlow, H. (1989). Unsupervised learning. Neural Computation 1, 295-311.

Bell, A. and Sejnowski, T., "An information-maximization approach to blind separation and blind deconvolution", *Neural Computation 7*, 1995, pp. 1129 - 1159.

Bishop, C., Svensen, M and Williams, C., "GTM: the generative topographic mapping", *Neural Computation 10*, 1998, pp. 215 - 234.

Cardoso, J.- F., "Blind signal separation: statistical principles", *Proc. of the IEEE 9 (10)*, 1998, pp. 2009 - 2025.

Grossberg, S, *Neural networks and natural intelligence*. Cambridge, MA: MIT Press, 1988.

Harman, H.H., *Modern Factor Analysis*. Univ. of Chicago Press, 1967.

Haykin, S., *Neural Networks - a Comprehensive Foundation*. New York: MacMillan College Publ. Co., 1998.

Hinton, G. and Sejnowski, T. (2000). Unsupervised Learning - Foundations of Neural Computation. MIT Press, Cambridge.

Hyvärinen, A., Karhunen, J., and Oja, E. (2001). Independent Component Analysis. Wiley-Interscience, New York.

Jutten, C. and Herault, J., "Blind separation of sources, part I: an adaptive algorithm based on neuromimetic architecture", *Signal Processing 24*, 1991, pp. 1 - 10.

Kohonen, T. (2001). The Self-Organizing Map. Springer, Berlin.

Kohonen, T., Kaski, S., Lagus, K., Salojärvi, J., Paatero, V. and saarela, A., "Self organization of massive document collection", *IEEE Trans. Neural Networks 11 (3)*, 2000, pp. 574 - 585.

von der Malsburg, C., "Self-organization of orientation sensitive cells in the striate cortex", *Kybernetik 14*, 1973, pp. 85 - 100.

NNRC (2001). Web pages of the Neural Networks Research Centre, Helsinki University of Technology. [Online reference, see <http:// www. cis.hut.fi/research/>, referred 20th July, 2001].

Oja, E., "A Simplified Neuron Model as a Principal Components Analyzer", *J. Math. Biol. 15*, 1982, pp. 267-273.

Oja, E., "Unsupervised learning in neural computation". *Theor. Comp. Science*, to appear (2001).
Oja, E. and Kaski, S. (Eds.),*Kohonen Maps*. Amsterdam: Elsevier, 1999.

Roweis, S. and Ghahramani, Z., "A unifying review of linear gaussian models", *Neural Computation 11 (2)*, 1999, pp. 305 - 346.

Tipping, M. E. and Bishop, C. M., "Mixtures of probabilistic principal component analyzers", *Neural Computation 11 (2)*, 1999, pp. 443 - 482.

Xu, L., "Temporal BYY learning for state space approach, hidden Markov model, and blind source separation", *IEEE Trans. Signal Proc. 48*, 2000, pp. 2132 - 2144.