

Inference and Learning with Graphical Models

Hideki ASOH, Yoichi MOTOMURA, Kôiti HASIDA
 Kazuhisa NIKI, Shotaro AKAHO, Masaru TANAKA
 Tatsuya NIWA, Kenji FUKUMIZU

Intelligent systems which work in the real world cooperating with humans should be able to treat complex, uncertain, dynamic, multi-modal information in the real world environment. Probability distributions with graphical structure (Graphical Models) are recently considered as a promising fundamental tools to represent and process such information. From this point of view, toward a theoretical and algorithmic foundation for real world intelligence, we are investigating inference and learning algorithms for several kinds of graphical models such as Bayesian networks, probabilistic constraint program, mixture models, multivariate models. Novel learning frameworks suited for real world intelligence are also investigated. Here we introduce some remarkable results from our recent researches.

§1 Introduction

What is the theoretical foundation of intelligent systems? About ten years ago, the strongest candidate was *symbolic logic*. Logic programming (declarative programming, constraint programming) was considered as a very powerful tool to represent knowledge about the world and to make inference with it. Nowadays logic is still an important part of the foundations but is considered not enough to handle complex, uncertain, dynamic, multi-modal information in the real world.

A linear architecture of intelligence which has complete knowledge about the environment and always makes correct inferences combining the knowledge and input information, then makes decisions and actions, is not supported now. Instead, more dynamical, situated, distributed architecture of intelligent systems have been proposed. There, an intelligent system is considered as a collection of many situated behavior-based agents which have incomplete, behavior specific knowledge and work making very close dynamic coupling loop including environment and the system itself.

As theoretical foundations for such situated agents,

two candidates are becoming major and are much more investigated recently. One is *dynamical systems* and the other is *probability models*. Same as symbolic logic, both of them have long history as languages for describing the world and making inference about the world. We do not think that those three are mutually exclusive. Instead, we are seeking for a good blend of them. In **Fig.1** some models used in intelligent system research are depicted with various kinds of information

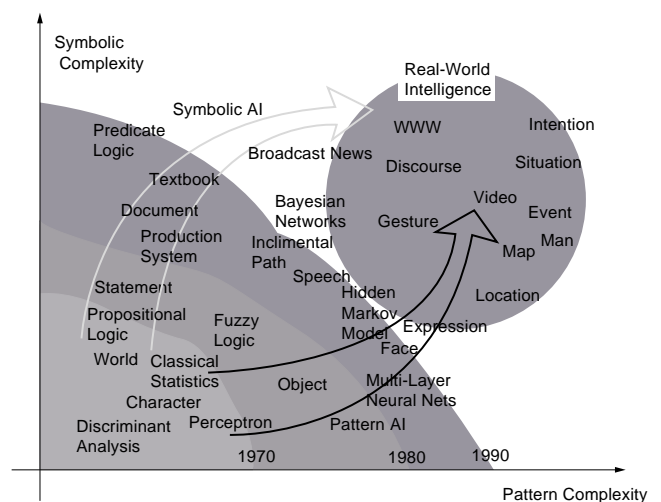


Fig.1 Real World Information and Models

in the real world. The base of our investigation is on the probability models. In particular, probability distributions with graphical structure are our starting point. As is well known, probability models were born as a theory of gambling, became to be used widely in statistics, and made great successes to acquire, represent and process statistical knowledge extracted from data. They are also very popular in pattern recognition and become a foundation of current character recognition and speech recognition systems that are recently commercialized.

In the history of the probability models, various families of models are proposed and applied to various kinds of data. However current target is still limited to single modal, well controlled data. In order to establish a theoretical foundation of real world intelligence, we should extend models to be able to handle multi-modal, more complex, more structured, not well controlled, non-stationary data.

From the point of view we are investigating inference and learning algorithms for several kinds of graphical probability distributions such as Bayesian networks, probabilistic constraint program, mixture models, multivariate models.

Adding to the inference and learning algorithms, we think that framework of learning is another very important research issue. In many conventional machine learning or statistics research, data acquisition, learning, and using the learned result are often separated. However, real world intelligence is expected to work in the real world having close coupling with the environment and it is not desired for them to separate the three stages of learning. More dynamic learning framework using the close coupling is needed and it may help to solve problems caused by shortage of training data or uncontrolled noisy training data.

In this paper we will introduce some remarkable results from our recent researches. In the next section, algorithms for graphical models are described. Novel learning frameworks are introduced in section 3, and in section 4 researches on optimization using probability models are described.

§2 Algorithms for Graphical Models

2.1 Bayesian Networks on Neural Networks

A Bayesian network consists of a directed acyclic graph and conditional probability distributions. Each node of the graph represents a random variable and a directed link represents conditional dependency (direct causal relation) between two variables connected by the link. A node which sends a link to another node is called a parent node of the target node. A conditional probability distribution of a random variable is assigned to the corresponding node, and conditioning part of the distribution is composed of random variables represented by parent nodes of the node.

The whole structure defines a joint probability distribution of random variables represented by nodes. If all variables are statistically dependent, then the graph becomes fully connected and nothing is different from usual multinomial distributions. However in many realistic cases a random variable is conditionally independent to many of other variables, and the graph becomes rather sparse. In such cases, by representing the dependence structure explicitly with a graph and using the structure to control the computation process efficient inference process, that is, computation of probability distribution of variables under a specific condition, can be realized.

Bayesian network can be considered as a generalization of major graphical models such as Hidden Markov models and finite mixture models and is becoming a popular tools for treating incomplete and noisy information in the real world.

One research issue is how to represent a conditional probability distribution assigned to each node. Normally a simple table is used. However when the number of conditioning variables increases, the size of the table increases exponentially. In addition, if some of conditioning variables takes continuous value, table is not usable to represent the distribution any more.

We proposed to use a neural network to represent a conditional probability of each Bayesian network node¹⁾. This idea leads to compact representation and efficient



Fig.2 GUI Snapshot of BAYONET³⁾

approximative probability computation. Learning the conditional probabilities from training examples is realized by learning algorithms for neural networks. Continuous valued variables can be treated also by the proposed model. We applied the model to several experimental applications such as prediction of weather data, context dependent handwritten character recognition, and localization of mobile robots to demonstrate the effectiveness of the model for situated probabilistic inferences²⁾.

We implemented a software environment for using our Bayesian network model easily. The software is written in Java and has a smart graphical user interface to build and edit the network structure, to run the inference procedure, and to acquire conditional probabilities from training examples [Fig.2]³⁾. In addition, it also has rich functions to connect to standard SQL databases and to extract training data from them. This is one of the first implementation of Bayesian Network in Java in the world and is listed in the Java-Repository.

2.2 Probabilistic Constraint Program

Bayesian networks use causal structures (constraints) represented by predicate logic sentences such as "if (A and B) then C". Extending this to first order logic is an interesting research issue. On the other hand, there is a tradition of logic programming and constraint-based

$$\begin{aligned}
 &\leftarrow P(A) \wedge q(A). \\
 p(X) &\leftarrow X = f(a) : 0.3. \\
 q(Y) &\leftarrow Y = f(b) : 0.7. \\
 q(Z) &\leftarrow Z = f(U) \wedge r(U) : 0.6. \\
 q(W) &\leftarrow Z = g(V) \wedge r(V) : 0.4. \\
 r(a) &: 0.8. \\
 r(b) &: 0.2.
 \end{aligned}$$

Fig.3 Probabilistic Constraint Program

programming (or declarative programming) to represent constraints in first order logic sentences. In view of this tradition, the research issue concerns an extension of logic program with probability.

Probabilistic Constraint Programming (PCP) is an attempt for an optimal combination of constraint logic program and probability theory. Its major feature is integration of symbol manipulation and probabilistic computation. A program (constraint) is a set of Horn clauses and a probability parameter is assigned to each clause [Fig.3], which derives a probability distribution over the program trees (candidates for proof trees).

Computational process is constraint satisfaction by program transformation to discard inconsistent program trees and obtain valid proof trees, seeking for probable solutions (sets of value bindings). Learning probability values from examples is possible with maximum entropy estimation. Introduction of symbolic learning technique like inductive logic programming is also within the scope.

This representational and computational framework has very strong power of describing and processing complicated constraints such as grammars of natural languages⁴⁾. The major concern here is how to avoid a huge computational cost for making inference and learning. Sophisticated control of computation is indispensable. We have introduced several techniques such as structure sharing, using memo to avoid redundant computation, and a new general compilation method which subsumes some standard ones⁵⁾.

2.3 Multivariate Information Analysis

In the context of Bayesian networks or probabilistic constraint program, constraints are in principle written

or designed by users. Logic based intuitive representation of the constraints is good for the purpose. However, to find out the relationship between variables is also an important research issue. Historically this issue is investigated in the domain of multivariate data analysis such as causal analysis. Recently researchers in Bayesian network community and KDD (knowledge discovery from data) community have interest in finding out causality from data.

Multivariate information analysis⁶⁾ includes methods to find out information exchange or sharing between random variables by analyzing mutual entropy between the variables. We formulate the method and applied it to analyze the data from functional MRI during experiments on cognitive functions of humans. fMRI data is composed of huge number of time series each of which represents transient of activities in a very small part of brain (boxel). By using the method we could find out relationships between boxels.

2.4 Mixture of Gaussians

Mixture of finite number of probability distributions is one of the simplest graphical models. In particular, mixture of Gaussian distributions is an important target for theoretical analysis. We investigated learning dynamics of the mixture of multiple Gaussians and clarify the structure of learning curve around the bifurcation points.

Theories of learning and generalization tell us that the generalization bias, which is defined as the difference between the training error and the generalization error, increases on average with the number of adaptive parameters. However, Akaho and Kappen has shown that this general tendency is violated for a Gaussian mixture model. For temperatures just below the first symmetry breaking point, the effective number of adaptive parameters increases and the generalization bias decreases [Fig.4]⁷⁾. We also applied mixture models and the EM algorithm to multiple object recognition⁸⁾.

2.5 Mixture of Predictors

Mixture models is normally used to segment input

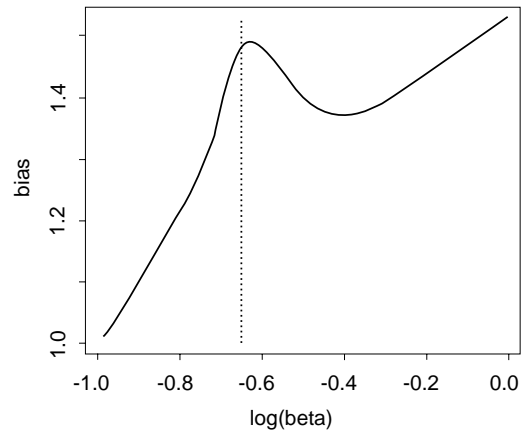


Fig.4 Generalization Bias of Mixture of Gaussians⁷⁾

signal space to several number of classes. However almost all of real world data such as speech waveform or motion pictures are multi dimensional time series. For such data, not only segmenting the signal space but also segmenting the data along time axis is an important research issue. The idea of using mixture of competitive dynamical predictors to segment time series data is a promising way to solve the problem. We investigated the property of the mixture of Elman networks using various kinds of data such as data from mobile robot sensors, motion picture sequences from RWC multi modal data base, and artificial chaotic time series and clarified the relation between parameters of networks and convergence of learning⁹⁾.

2.6 Neural Network Models

Neural network model is normally used as a pattern classifier or a function approximator. However it is also possible to consider learning of neural network as estimation of a probability distribution, which generates the statistical training data. Here, connection weights are parameters of the distribution and learning of the network is estimating the optimal parameter values. Maximum likelihood estimation or Bayesian estimation of parameters are investigated.

In this framework, a specific character of the multi-layer neural network model is its powerful description capability. As is well known, any continuous function can be approximated by a three layer (one hidden layer) neural network with a proper number of hidden units.

A drawback of the neural network model is the possibility that Fisher information matrix at the true parameter may become singular. It may happen something special when such singularity is satisfied. This singularity always occurs if we consider the model selection problem in neural networks. The limiting distribution of the MLE is not obtained by the ordinary statistical asymptotic theory in this case, and any methods based on the asymptotic theory such as AIC and MDL are not applied properly.

We evaluated generalization error of linear neural network model as the simplest multi-layer model, and explicitly show that the generalization error becomes larger when the Fisher information matrix at the true parameter value is singular than the generalization error derived from the usual asymptotic theory¹⁰⁾. In addition, while the ordinary asymptotic theory asserts that the expectation of the generalization error depends only on the number of parameters, the generalization error in linear neural networks depends on the rank of target function also.

§3 Learning Framework

As we mentioned in the introduction, conventional machine learning researches are mainly investigating the learning capability within the framework of "learning from examples." There it is assumed that a huge amount of training data are prepared and learning is executed off-line. The phase of training and the phase of task execution is separated.

To the contrary, the real world intelligence systems should treat various multi-modal information from sensor data to symbolic language in a dynamic environment. Here it is not expected to have enough training data beforehand. The systems are requested to gathering the training data during task execution by themselves. In order to take the learning system to the real world, we should solve many problems such as learning from small amount of data, learning from uncontrolled, not well organized data, etc.

Adding to the model and algorithms, we are investigating

new framework of learning to solve the problems. Here we introduce two major results, *dialog-based learning* and *multi-attribute learning*.

3.1 Dialog-based Learning

A way to solve the problem of small number of training data is using additional information such as heuristics, hints, etc. The explanation-based learning is an example of such learning where domain knowledge represented by rules is used to assist learning processes. Another kind of additional information comes from dense interaction with environments during learning. For example, human children apparently have very dense interaction with their surroundings, especially with their parents and it helps the children's learning process very much.

A very powerful communication channel between learning systems and human users is dialogue with natural or semi-natural language. Here is a question: "How effectively the dialogue between systems and users can be used in learning process or can help the system to learn ?"

The idea of "dialogue-based learning" which exploits dialogue in natural language for teaching is rather old. Although the idea is very simple and natural, the necessity of the capability of speech understanding has been a bottleneck and not so much effort to realize a system has been made. Recently, as a result of the progress of AI and pattern recognition research, the technique of speaker independent continuous speech recognition and natural language understanding has reached to an applicable level, and the dialogue-based learning are becoming attractive again for building real-world oriented autonomous learning systems.

We applied this idea to the topological map learning problem in the autonomous mobile robot navigation. The implemented system on a real mobile robot *Jijo-2* can make simple spoken dialogue with a human trainer and succeeded to acquire a graphical map of the environment^{11,12)}. We are now extending the target to more general models of environment including various kinds of maps, task models, and user models.

3.2 Multiple-Attribute Learning

Interactions between users and intelligent systems should not be strongly controlled. This implies that the learning processes of the systems should not be controlled strongly also. Let us imagine that you will teach a system about your favorite mug cup. The mug cup may be white large one. You'll show the cup to the system and may tell to the agent that "This is white." Or you may tell to the system that "This is large." The visual image of a mug cup has at least three attributes, color, size, and shape. Hence there are at least three ways of classification of a visual image, classification by color, by size, and by shape.

In a normal setting of pattern recognition, one way of classification is chosen in advance. However, in the real situation of concept learning of interactive systems, learning data for multiple classification problems may be given to the system at once. The system should treat such uncontrolled data set and learn multiple classification rules from the data simultaneously.

Note that the system does not know about what kinds of attribute the input signals have and will discover the attributes such as color, size, shape during learning. Once the system discover the attribute, adding new category to the attribute, for example, adding "pink" to color attribute, may become easier task. We investigated this problem of "multiple attribute learning" and proposed learning algorithm based on the EM algorithm¹³⁾.

§4 Optimization

As is pointed out by some researchers, optimization is closely related to probabilistic inference and learning. Let us consider that an unknown blackbox function is given and you are requested to find the maximum point of the function. If the target function is written as a continuous differentiable function, gradient based methods can be used. However, in the real world the target function may be a very complicated black box. In such cases only you can execute is to input some sample points to the function and observe outputs. Using

the information you will revise your guess about the location of the maximum point. This process is very similar to an active learning process.

To represent the guess with probability distribution is a natural idea. According to the distribution next sample points will be generated. Repeating this two processes, generating population of sample points and revising the distribution, leads to population based random search methods like Genetic Algorithms. Here two main research issues are how to select sample points and how to revise the guess using the observations at sample points. From this point of view we are investigating dynamics of some optimization methods.

4.1 Genetic Algorithms

Genetic algorithms are optimization methods inspired by the process of biological evolution, and have been succeeding to solve difficult optimization problems. However, the dynamics of GAs to solve problems is complex and has not been understood thoroughly.

To understand the behavior of GAs, we started to analyze dynamics of a very simple and typical genetic process which is called "Wright-Fisher model" in the field of population genetics. We evaluated the mean convergence time of the model with genetic drift because of the finite population size both theoretically and experimentally, and found a reference value of the most efficient mutation rate for canonical GAs¹⁴⁾.

Then we extended the model to one of the parallel model, which is called "island model". In the model, the global population is divided into several subpopulations and migration between them is introduced. We evaluated the mean convergence time with genetic drift in the model and derived a critical value of the efficient migration rate¹⁵⁾.

4.2 Population Search

So far some optimization algorithms using probability distributions to model the guess about the optimum point has been proposed. They use several kinds of statistical models such as special form of Bayesian networks, and the algorithms show superior performance to other

random algorithms.

For the sake of simplicity of analysis we formalized an algorithm to use a Gaussian distribution to model the guess and to use truncate selection to revise the guess. We analyzed the convergence property of the optimization process theoretically and found that the algorithm based on the explicit modeling of the elites' distribution tends to converge to local optima. We also proposed a modified algorithm to overcome the defect¹⁶⁾.

§5 Conclusion

Recent results from our researches are described. Toward establishing a theoretical and algorithmic foundation of real world intelligence, we have been conducting wide variety of researches from theoretical analysis to developing practical algorithms and implementing software. Various kinds of graphical models and learning frameworks are studied in the researches.

We will apply our practical algorithms and learning frameworks to the prototypes of real world intelligence such as office robot Jijo-2 and multi-modal interaction system, and evaluate the effectiveness. At the same time we will continue theoretical researches to find interesting properties of the models and algorithms.

Acknowledgements

This work is done as a part of the Real-World Computing Project conducted by RWCP and ETL under supervision of the MITI. The authors greatly thank all persons who are managing the project.

References

1) Y. Motomura, I. Hara, H. Asoh, and T. Matsui, "Bayesian network that learns conditional probabilities by neural networks", *Proc. of 4th Int. Conf. on Neural Information Processing* (1997) 584-587.

2) Y. Motomura, "Integration of situated prior probability and neural network classifier in a handwritten recognition task",

Proc. of the 5th Int. Conf. on Neural Information Processing (1998) 283-286.

3) Y. Motomura, "BAYONET: Probabilistic reasoning system with learning from database", *Proc. of the 12th Annual Conf. of JSAI 1998* (1998) 632-633 (in Japanese).

4) K. Hasida, "Constraint-based derivation of cognitive model on parsing", *Proc. of 1997 Int. Conf. on Cognitive Science* (1997) 51-56.

5) K. Hasida, "Parsing and Generation with Tabulation and Compilation", *Proc. of TAPD '98* (1998) 26-35.

6) K. Niki, J. Hatou, and I. Tahara, "Structure analysis and activity mapping of fMRI by using multivariable information", *Proc. of 1999 Annual Conf. of JNNS* (1999) 140-141 (in Japanese).

7) S. Akaho and B. Kappen, "Nonmonotonic generalization bias of Gaussian mixture models", *Neural Computation*, 2000 (in press).

8) S. Akaho, "The EM algorithm for multiple object recognition", *Proc. of 1995 IEEE Int. Conf. on Neural Networks* (1995) 2426-2431.

9) K. Horikawa, H. Asoh, J. Tani, T. Matsui, and M. Kakikura, "Emergence of expert modules for mobile robot navigation from a mixture of Elman networks", *Proc. of 5th Int. Conf. on Neural Information Processing* (1998) 256-259.

10) K. Fukumizu, "Generalization error of linear neural networks in unidentifiable cases", *Proc. of 10th Int. Conf. on Arithmetic Learning Theory* (1999).

11) H. Asoh, Y. Motomura, I. Hara, S. Akaho, S. Hayamizu, and T. Matsui, "Combining probabilistic map and dialog for robust life-long office navigation", *Proc. of 1996 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems* (1996) 807-812.

12) H. Asoh, S. Hayamizu, I. Hara, Y. Motomura, S. Akaho, and T. Matsui, "Socially embedded learning of the office-conversant robot Jijo-2", *Proc. of 15th Int. Joint Conf. on Artificial Intelligence* (1997) 888-885.

13) S. Akaho, S. Hayamizu, O. Hasegawa, T. Yoshimura, and H. Asoh, "Concept acquisition from multiple information sources by the EM algorithm", *Trans. IEICE J80-A* (1997) 1546-1553 (in Japanese).

14) T. Niwa and M. Tanaka, "On the mean convergence time for simple genetic algorithms", *Proc. of 1995 Int. Conf. on Evolutionary Computing* (1995).

15) T. Niwa and M. Tanaka, "Analysis on the island model parallel

genetic algorithms for the genetic drifts”, *Proc. of 2nd Asia-Pacific Conference on Simulated Evolution and Learning* (1998).

- 16) S. Akaho, “Statistical learning in optimization: Gaussian modeling for population search”, *Proc. of 5th Int. Conf. on Neural Information Processing* (1998) 675-678.

(Accepted May 1, 2000)