Special Feature on Discovery Science

Hiroshi MOTODA

The Institute of Scientific and Industrial Research, Osaka University, 8-1 Mihogaoka, Ibaraki, Osaka 567-0047, JAPAN

Setsuo ARIKAWA

Department of Informatics, Kyushu University, Hakozaki 6-10-1, Fukuoka 812-8581, JAPAN

motoda@sanken.osaka-u.ac.jp, arikawa@i.kyushu-u.ac.jp

Received 30 August 1999

§1 Introduction

Discovery has always attracted people and has been a major motive force of progress in human history. It has long been discussed in the philosophy of science but it is only after 1960's that the problem of discovery became properly addressed by new philosophy of science group [6]. In the last two decades, there has also been a considerable progress, in the field of artificial intelligence, of a domain often considered the realm of genius - empirical discovery [5]. Historically, discovery tends to be viewed as a difficult form of learning from observation although learning and discovery convey rather different meanings [4]. The former suggests a gradual process, while the latter suggests a more rapid mental event, often involving some form of insight. Learning may lead to an unconscious change in knowledge, while one is always aware that a discovery has been made. The results of learning can be declarative or procedural, while the product of discovery is always declarative. Despite these differences, the algorithms and the methods used in discovery research owe much the results of machine learning research. Discovery is also a part of knowledge acquisition although much of the work in this field has been to acquire and operationalize knowledge from human

experts. Another important field closely related to discovery is data-mining that has emerged very rapidly in the last decade. The advent of the age of digital information has brought the problem of data overload. Our ability to analyze and understand massive datasets lags far behind our ability to gather and store the data, and a new generation of computational techniques and tools is required to support the acquisition of useful knowledge from the rapidly growing volume of data. In addition to these, statistics without saying has always been used in analyzing data and extracting useful knowledge from them. In view of these all, we believe that the time is ripe to integrate all of these techniques that have been developed more or less in isolation into a new paradigm: discovery science.

A new project "Discovery Science" [3] has been launched last year that targets to 1) develop new methods for knowledge discovery, 2) install network environments for knowledge discovery, and 3) establish the Discovery Science as a new area of computer science, under the auspices of Grant-in-Aid for Scientific Research on Priority Area from the Ministry of Education, Science, Sports and Culture of Japan. A systematic research is being conducted that ranges over philosophy, logic, reasoning, computational learning and system developments. The first international conference on discovery science (DS98) sponsored by this project was held in December, 1998 [2]. The second one DS99 will be held in Decemeber, 1999.

It is very timely to have a special feature of discovery science in this issue. The papers collected in this feature were selected from among those presented in DS98 and have been updated to include recent advancement. All of them are grounded on sound theory, challenge new dimensions of discovery science, and have demonstrated their applicability to practical yet hard problems.

§2 Papers in this Special Feature

The first paper by Kitagawa and Higuchi shows the power of statistical data analysis method using AIC. Log likelihood is known as an estimate of the Kullback-Leiber information which is a measure of similarity between the predictive distribution of the model and the true distribution. Noting that AIC is an estimate of the K-L information, they used this measure to find the best model. This is a breakthrough in statistics and changed the paradigm from estimating parameters within a given structure to selecting the best model from different structures. They showed that this techniques can indeed discover a very subtle change in the underground water level due to an earthquake from

very noisy data.

The second paper by Pazzani proposes a method to improve the understandability of learned rules by introducing new biases. Rules created by existing rule learning systems are not mutually exclusive. They are ordered and the first rule that fires is used to classify an example. Thus, each rule must be interpreted under the context it is used, and it is possible that the rules contain certain tests that are counter-intuitive and puzzling to expert if they are interpreted in isolation. Pazzani's method avoids this from happening. His first bias is to only allow to create a rule that has globally predictive tests (*i.e.*, $P(Class_i|Test) > P(Class_i)$). The second one is to relax the first one to prefer the globally predictive tests to locally predictive ones unless the latter is statistically significant. Interestingly, these biases sometime increase the accuracy and suggest that they may aid in preventing overfitting. They represent a form of simplicity bias and are useful to avoid overly complex models when simpler explanations of the data are possible.

The third paper by Shinohara, et. al. challenges the task of retrieving similar objects from a huge number of high-dimensional spatial data. The distance metric they used is discrete L_1 (Manhattan distance) which is very general. The key idea is to use the spatial indexing of R-tree by projecting the objects into a space of low dimension using FastMap that assumes Euclidean distance. They solved this problem by finding that taking the square root of L_1 distance enables the objects to be embedded in a Euclidean space. They applied their method to a problem of retrieving Japanese chess boards that are similar to the one given in a query. The number of candidates boards are 40,000 and their original dimension is 2,300. They reduced the dimension to about 10 and obtained surprisingly good results. This method will be useful to data mining of such high dimensional data as documents, digital images and audio clips.

The fourth paper by Shimozono et. al. aims at discovering optimal word association patterns in large text databases. Their aim is to provide an efficient tool for data mining that can be used for weakly structured data such as bibliographic databases, e-mails, HTML statements and raw experimental data (*e.g.*, genomic sequences). They developed an algorithm that can find a k-proximity d-word association pattern that maximizes the number of texts labeled positive (positive documents) matching this pattern. The algorithm finds the best pattern in time and space almost linear to the total text input length n (time complexity $O(k^{d-1}n\log^d n)$ and space complexity $O(k^{d-1}n)$), which is

drastically faster than a straight-forward algorithm of $O(n^{2d+1})$ time complexity. The paper focuses on the theoretical aspect of this algorithm, but it has already been implemented and applied to Reuters-21578 text database and GenBank database [1)] with good performance as expected.

The fifth paper by Yamasaki et. al. attempts to characterize a form of traditional Japanese poetry called Waka using a pattern discovery technique. Waka, which has a history of 1,300 years, was used as a subtle means of communication, and was usually composed in momentary flashes of inspiration. The pattern they focused is a particular form of expression called Fushi composed of adjuncts, which is thought of a reflection of writer's personality and is a rhetorical device. Fushi pattern is represented as a regular expression. Since there are many patterns that are covered by different Waka poems, it is important to define their significance. They used MDL principle to prune the patterns. What is interesting is that this measure is better than the well-used impurity measure, and they were able to characterize the Waka poems from different anthologies. They also found Fushi patterns that are non-obvious. It is interesting to see the challenge of a machine discovery method to a problem that needs deep semantic understanding.

The sixth paper by Ohsawa and Yachida proposes an interesting approach that can predict the existence of unknown causes by analyzing the unexpected co-occurence of known events. The events that were inferred to have occurred simultaneously were identified by cooperative abductive inference. It is based on the belief in coherency, *i.e.*, a hypothesis in a state that also holds in its adjacent states is more likely to be true than those that are true in only that state. Multiple abducers exchange messages which in turn change their own beliefs and converge to an agreement. They showed that this method can infer such unknown causes as "Two lines touched intermittently" when applied to a circuit diagnosis problem.

References

1) Arimura, H., Wataki, A., Fujino, R. and Arikawa, S., "A fast algorithm for discovering optimal string patterns in large text databases," In Proc. the 9th Int. Workshop on Algorithmic Learning Theory, LNAI 1501, Springer-Verlag, pp.247–261, 1998.

2) Arikawa, S. and Motoda, H. (Eds.), Discovery Science, *First International Conference, DS'98, Lecture Notes in Artificial Intelligence*, LNAI 1532, Springer-Verlag (1998)

3) "http://www.i.kyushu-u.ac.jp/~arikawa/discovery/indexE.html".

4) Langley.P., "Editorial: Machine Learning and Discovery," *Machine Learning*, 1, pp.363–366, (1986)

5) Langley.P., "Data-Driven Approaches to Empirical Discovery," *Artificial Intelligence*, 40, pp.283–312, (1989)

6) Noe, K., "Philosophical Aspect of Scientific Discovery: A Historical Survey," *First International Conference, DS'98, Lecture Notes in Artificial Intelligence*, LNAI 1532, Springer-Verlag, pp.1–11, (1998)