Theoretical Frameworks for Data Mining

Heikki Mannila
Nokia Research Center
P.O. Box 407 (Itamerenkatu 11)
FIN-00045 Nokia Group, Finland
Heikki.Mannila@nokia.com

1. INTRODUCTION

Research in data mining and knowledge discovery in databases has mostly concentrated on developing good algorithms for various data mining tasks (see for example the recent proceedings of KDD conferences). Some parts of the research effort have gone to investigating data mining process, user interface issues, database topics, or visualization [7]. Relatively little has been published about the theoretical foundations of data mining. In this paper I present some possible theoretical approaches to data mining. The area is at its infancy, and there probably are more questions than answers in this paper.

First of all one has to answer questions such as "Why look for a theory of data mining? Data mining is an applied area, why should we care about having a theory for it?" Probably the simplest answer is to recall the development of the area of relational databases. Databases existed already in the 1960s, but the field was considered to be a murky backwater of different applications without any clear structure and without any interesting theoretical issues. Codd's relational model was a nice and simple framework for specifying the structure of data and the operations to be performed on it. The mathematical elegance of the relational model made it possible to develop advanced methods of query optimization and transactions, and these in turn made efficient general purpose database management systems possible. The relational model is a clear example of how theory in computer science has transformed an area from a hodgepodge of unconnected methods to an interesting and understandable whole, and at the same time enabled an area of industry.

Given that theory is useful, what would be the properties that a theoretical framework should satisfy in order that it could be called a theory for data mining? The example of relational model can serve us also here. First of all, the theoretical framework should be simple and easy to apply; it should (at least some day) gives us useful results that we could apply to the development of data mining algorithms and methods.

A theoretical framework should also be able to model typical data mining tasks (clustering, rule discovery, classification), be able to discuss the probabilistic nature of the discovered patterns and models, be able to talk about data and inductive generalizations of the data, and accept the presence of different forms of data (relational data, sequences, text, web). Also, the framework should recognize that data mining is an interactive and iterative process, where comprehensibility of the discovered knowledge is important and where the user has to be in the loop, and that there is not a single criterion for what an interesting discovery is. (We could also ask, "What actually is a theory?" For that I have the simple answer: we recognize a theory when we see it.)

I start by discussing reductionist approaches, i.e., ways of looking at data mining as a part of some existing area, such as statistics or machine learning; in this case, of course, there is little need for new theoretical frameworks. Then I discuss the probabilistic approach, which is of course closely linked to statistics: it views data mining as the activity aimed at understanding the underlying joint distribution of the data. After that, I review the data compression approach to the theory of data mining. The very interesting microeconomic viewpoint on data mining is considered after that, and finally I look at the concept of inductive databases, and show how it can perhaps be used to understand and develop data mining.

2. TWO SIMPLE APPROACHES

A simple approach to the theory of data mining is to declare that data mining is statistics (perhaps on larger data sets than previously), and thus the search for a theoretical framework for data mining can stop immediately: we just have to look at the appropriate statistics literature. The theory of data mining is statistics (as a science).

Data mining obviously is very close to statistics, and data mining researchers with computer science backgrounds (including myself) typically have too little education in statistics. However, one can argue that there are important differences between the areas. The volume of the data is probably not a very important difference: the number of variables or attributes often has a much more profound impact on the applicable analysis methods. For example, data mining has tackled with problems such as what to do in situations where the number of variables is so large that looking at all pairs of variables is computationally infeasible. Overall, the issue of computational feasibility is has a much clearer role in data mining than in statistics. Another difference pointed out to me by David Hand is that data mining is typically secondary data analysis: the data has been collected for some other purpose that for answering a specific data analytical question.

Several other differences between the areas could be pointed out. For example, the emphasis on database integration, simplicity of use, and the understandability of results are typical of data mining approaches. For the purposes of this paper it is sufficient to point out that at least currently the theoretical framework of statistics seems to be relatively distant from the actual development of data mining methods. Also, statistical theory does not seem to pay a lot of attention to the process character of data mining. However, in the next section we describe a closely related approach.

A similar (but weaker) case of reducing data mining to an existing area has been made from the viewpoint of machine learning. One could say that data mining is applied machine learning, and thus the theory of data mining is equal to the theory of machine learning. Again, this approach fails for two reasons. First, there are important differences between machine learning and statistics, and second, the theoretical machine learning approaches (such as the PAC-model) do not really address the special requirements that we made for the theory of data mining.

3. PROBABILISTIC APPROACH

A possible theoretical approach to data mining is to view data mining as the task of finding the underlying joint distribution of the variables in the data. Typically one aims at finding a short and understandable representation of the joint distribution, e.g., Bayesian network [10] or a hierarchical Bayesian model [8,9].

This approach is obviously closely related to the reductionist approach of viewing data mining as statistics. The advantages of the approach are that the background is very solid, and it is easy to pose formal questions. Tasks such as clustering or classification fit easily into this approach. What seems to be lacking, as in most of the approaches, are ways for taking the iterative and interactive nature of the data mining process into account.

Hierarchical Bayesian models [8,9] seem a very promising statistically sound approach to data mining. Such a model describes the structural part of the distribution independently of the actual functional form of the distribution. For example, if we have information about the supermarket-buying behavior of a group of people, we could describe the model using the diagram in the figure below. That is, for each customer there is a set of visits to the shop, and for each visit of a customer there is a set of products bought during that visit. To give a probabilistic model for this phenomenon we have define how the number of visits is distributed for each customer, and how the selection of products is distributed for each visit. What is interesting to notice is that the figure could be interpreted either as part of the probabilistic model or just simply as an ER schema for the data set. When statisticians and database people talk about data modeling, they are talking about different things, but there still is lots of commonality in the meanings. For hierarchical models such as the one shown below there exists good tools for approximating the posterior distribution of a model [8,9]. It seems to me that exploring the usefulness of such models for data mining is a promising area of research: issues such as scalability etc. need to be addressed.

4. DATA COMPRESSION APPROACH

The data compression approach to data mining is simply to state: the goal of data mining is to compress the data set by finding some structure for it. That is, data mining looks for knowledge, where knowledge is interpreted as a representation that makes it possible to code the data using few bits. If desired, the minimum description length (MDL) (see, e.g., [15]) principle can be used to select among different encodings. To yield structure that is comprehensible to the user, we have to specify compression methods that are based on concepts that are easy to understand.

Several simple data mining techniques can be viewed as instances of this approach. For example, association rules [1,2] can be viewed as ways of providing compression of parts of the data. In the same way, an accurate decision tree can be considered a compression method for the target attribute. A clustering of the data can also be viewed as a way of compressing the data set.

This approach has a nice formal foundation. It is connected to Bayesian approaches for modeling the joint distribution: any compression scheme can be viewed as providing a distribution on the set of possible instances of the data. As in the probabilistic approach, the process view of data mining is not so easy to capture in this framework. An interesting opening in this direction is, however, given in [5], where it is shown how to mine for novel nuggets using the MDL principle.

5. MICROECONOMIC VIEW OF DATA MINING

The microeconomic view of data mining introduced by [13] is a very interesting approach. The starting point is that data mining is about finding actionable patterns: the only interest is in patterns that can somehow be used to increase utility. Kleinberg et al. give a decision theoretic formulation of this principle: the goal of the organization is to find the decision x that leads to the maximum utility $f(x)$. The form of the utility $f(x)$ is typically a sum of utilities $f_i(x)$ for each customer i. This function $f_i(x)$ is actually a complex function of the decision x and the data y_i on customer i, and can often be represented using a single function, i.e., as $f_i(x)= g(x, y_i)$. Thus the task is to find the decision x maximizing the sum of the terms $g(x, y_i)$ over the customers i. The basic observation of Kleinberg et al. is that data mining is useful if and only if the function g is nonlinear. They are able to describe pattern discovery, clustering, etc. as instantiations of the framework, and demonstrate also interesting connections to sensitivity analysis. Furthermore, they also show how the framework gives useful suggestions for research problems. I cannot do justice to this delightful paper in this article, but the approach is clearly very promising.

6. INDUCTIVE DATABASES

As mentioned above, relational database theory has been remarkably successful. One of the basic concepts in that theory is the powerful notion of a query. Instead of thinking of accessing the database as a special entity, relational database theory just considered queries as functions mapping databases to databases. This made it possible to speak about composing queries etc., and in many ways was instrumental to the development of relational databases.

The basic idea of inductive databases is that the query concept should be applied also to data mining and knowledge discovery tasks. In the slogan form from [11, 12]: *there is no such thing as*

discovery, it is all in the power of the query language. That is, one can benefit from viewing the typical data mining tasks not as dynamic operations constructing new nuggets of information, but as operations unveiling hitherto unseen but pre-existing pieces of knowledge.

The term *inductive database* [14, 3, 4] refers to a normal database plus the set of all sentences from a specified class of sentences that are true of the data. In model-theoretic terms [6], the inductive database contains the data and the theory of the data.

The approach can be compared to the idea of deductive databases, which contain a normal database plus a set of rules for deriving new facts from the facts already existing in the database. The user of a deductive database can act as if all the facts derivable from the database would be actually stored there. Of course, this set might be infinite, or finite but very large, so in practice it cannot be represented. But the idea of treating stored and derived facts in the same way is crucial for deductive databases.

In the same way, an inductive database does not contain all the rules that are true about the data stored in it; the user is just able to assume that all these rules are there. In practice, the rules are constructed on demand. The schema of an inductive database consists of a normal relational database schema plus a schema for the generalizations. It is relatively easy to design a query language that works on such schemas [4]. The result of a query on an inductive database is again an inductive database, so we have the closure property that has been so useful for relational databases.

The process view on data mining is directly built in to the concept of inductive databases. It also suggests architecture for data mining systems. Association rules and other simple pattern formalism fit quite easily into the framework, and there are some good partial solutions that can be viewed as partial implementations of inductive databases. However, e.g., clustering is harder to describe in a useful way. The probabilistic nature of data mining can be incorporated by having the underlying concept class support probabilistic concepts.

7. CONCLUDING REMARKS

We required that a good theory for data mining should consider the process of data mining, have a probabilistic nature, be able to describe different data mining tasks, be able to allow for the presence of background knowledge, etc. None of the above candidates satisfies all the requirements, unfortunately.

My current favorite approach would be to combine the microeconomic view with inductive databases (or some other database oriented approach): these two aspects would seem to satisfy most of the requirements, and both directions suggest a wealth of interesting research issues.

8. REFERENCES

[1] R. Agrawal, T. Imielinski, and A. Swami: Mining Association Rules between Sets of Items in Large Databases, SIGMOD'93, p. 207-216.

[2] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A.I. Verkamo: Fast discovery of association rules, Advances in Knowledge Discovery and Data Mining", *Advances in Knowledge Discovery and Data Mining*, U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (eds.), AAAI Press 1996, p. 307-328.

[3] J.-F. Boulicaut, M. Klemettinen, and H. Mannila: Querying inductive databases: a case study on the MINE RULE operator. 2nd European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD'98)}, Nantes, France, September 23-26, 1998. p. 194-202.

[4] J.-F. Boulicaut, M. Klemettinen and H. Mannila: Modeling KDD Processes within the Inductive Database Framework. Data Warehousing and Knowledge Discovery (DaWaK 1999)}, M.K. Mohania and A. M. Tjoa (eds), p. 293-302

[5] S. Chakrabarti, S. Sarawagi, and B. Dom: Mining Surprising Patterns Using Temporal Description Length. *VLDB* 1998, p. 606-617

[6] C. C. Chang and H. J. Keisler: *Model Theory*, North-Holland 1973 (3rd ed., 1990).

[7] U.M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth: From Data Mining to Knowledge Discovery: An Overview. In *Advances in Knowledge Discovery and Data Mining*, U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (eds.), AAAI Press 1996, p. 1-34.

[8] A. Gelman, J. B. Carlin, H. S. Stern and D. B. Rubin: *Bayesian Data Analysis*, Chapman & Hall 1995.

[9] W.R. Gilks, S. Richardson, and D.J. Spiegelhalter: *Markov Chain Monte Carlo in Practice*, Chapman & Hall, 1996.

[10] D. Heckerman: Bayesian Networks for Data Mining. Data Mining and Knowledge Discovery 1, 1 (1997), 79-119.

[11] T. Imielinski: A database view on data mining. Invited talk at KDD-95.

[12] T. Imielinski and H. Mannila: A database perspective on knowledge discovery. *Communications of the ACM* **39**, 11 (1996), 58-64.

[13] J. Kleinberg, C. Papadimitriou, and P. Raghavan, A Micro-economic View of Data Mining, *Data Mining and Knowledge Discovery* **2**, 4 (1998), 311-324

[14] H. Mannila: Inductive databases and condensed representations: concepts for data mining. *International Logic Programming Symposium* 1997, p. 21-30.

[15] J. Rissanen: Hypothesis selection and testing by the MDL principle. *The Computer Journal* **42**, 4 (1999), 260-269.

About the author:

Heikki Mannila is Research Fellow at Nokia Research Center in Helsinki, Finland, where he leads a data mining research group. He is also a professor at the Department of Computer and Information Science of the Helsinki University of Technology. He has prveviously worked at Microsoft Research and the University of Helsinki. He is the editor-in-chief of *Data Mining and Knowledge Discovery*. His research interests include data mining, algorithms, and databases.