EVIDENCE BASED DISCOVERY OF KNOWLEDGE IN DATABASES

Sarabjot S. Anand, David A. Bell[1] and John G. Hughes[2]

Introduction

The amount of data being stored in databases has been on the increase since the 1970's partly due to the advances made in database technology since the introduction of the relational model for data by E.F. Codd [CODD70]. While the data storage and handling mechanisms have developed rapidly to cope with the increasing volume of data stored on computers, software tools for analysing this data and utilising it have been slow to develop and are far from satisfactory.

Recent advances in the field of machine learning [MICH83] have made the development of intelligent automatic analysis tools for the data a reality. But machine learning alone is not the answer to the problems faced in the real world in data analysis. Most machine learning algorithms get quite inefficient when it comes to using them with large quantities of data. Thus what is required is an amalgamation of machine learning and database expertise to develop good, efficient methods for data analysis.

Also, databases are designed for purposes other than discovery and therefore pose a number of problems within the discovery process. Data in databases is not static and methods for updating the knowledge discovered from databases are required to keep the discovered knowledge consistent with the data in the database. Databases often contain noise and missing values that need to be taken into account within the discovery process. Also, the enormous size of the data means that efficiency of the algorithms is essential. Another problem is integrating knowledge discovered from different databases and the incorporating of data from different types of databases within a single discovery process.

Databases also contain certain domain knowledge, e.g. integrity constraints, that can be incorporated within the discovery process to improve the efficiency of the process [BELL93]. Indices and other structures within databases may be used to improve the efficiency of the algorithms as well. Clearly, a number of new interesting problems and challenges arise from knowledge discovery in databases which were not faced by machine learning researchers and solutions to these problems can only be found by the integration of database and machine learning expertise.

Most work in database mining has been based around probabilistic models [PIAT91, PIAT93, FAYY94]. The authors have been investigating the use of Evidence Theory [GUAN91, GUAN92, SHAF76], a generalisation of the Bayesian Model for uncertainty, in Database Mining [ANAN94a]. We have proposed a general framework for Database Mining based on Evidence Theory [ANAN94]. The framework provides a common method for representing and manipulating data and knowledge in the form of generalised mass functions. The discovery process consists of a set of operations on these mass functions the result being the discovery of knowledge. The framework provides facilities that are common to all discovery processes e.g. methods for incorporating domain knowledge, dealing with missing values etc. The framework is inherently parallel and is, therefore, expected to be efficient for large data sets. The framework easily extends to discovery in parallel and distributed databases. Enhancing the framework to cope with discovery of different types of knowledge is simple as it means the addition of a few new operators to the framework.

[1] Sarabjot S. Anand, D. A. Bell are with the School of Information and Software Engineering, University of Ulster at Jordanstown, Northern Ireland

[2] John G. Hughes is with the Faculty of Informatics, University of Ulster at Jordanstown, Northern Ireland

The Bayesian Model for uncertainty is more a representation of the ideal reasoning technique rather than that used by humans thus making it a prescriptive rather than a descriptive method for reasoning. Thus, we feel that Evidence Theory has potential for use in the field of Database Mining [ANAN94a]. There are two main advantages of using an Evidence Theory-based approach over the Bayesian approach. Firstly, Evidence Theory provides mechanisms for coarsening and refining and thus for the combination of evidence at different levels of coarseness in a hierarchical space which clearly has uses in database mining. Secondly, it provides a method for representing *ignorance* which is an intuitive way of dealing with missing values in a databases during the discovery process.

The framework has been used to implement an algorithm for the discovery of strong rules and has also been shown to extend to the mining of knowledge in Spatial Databases [BELL94].

Evidence Theory in Knowledge Discovery

Evidence Theory is a generalization of the Bayesian Model for Uncertainty that allows the notion of partial belief. The basic probability assignment function (also known as the mass function) assigns belief to sets of propositions rather than just singletons. This is possible due to the relaxation of the Law of Additivity (3rd Kolmogorov Axiom) in Probability Theory.

There are two main advantages of using Evidence Theory as the model of uncertainty within Knowledge Discovery in Databases [ANAN94a]. Firstly, Evidence Theory allows a degree of Belief to be associated with Ignorance which is a natural way of considering missing values in databases. Secondly, Evidence Theory provides mechanisms for the combination of evidence at different levels of coarseness that leads to parallel algorithms for knowledge discovery. It allows for knowledge discovery from heterogeneous, parallel and distributed databases.

A Framework for Knowledge Discovery in Databases

The advantages of using Evidence Theory as the model of uncertainty for Knowledge Discovery in Databases prompted the authors to develop a framework for Database Mining/ Knowledge Discovery in Databases [ANAN94].

The framework consists of a method for representing data and knowledge and methods for knowledge discovery. Having such a general framework has a number of advantages. Firstly, the common method for representing knowledge allows the incorporation of prior knowledge from the user or that discovered by other discovery processes in the discovery of new knowledge. Secondly, the framework provides facilities that are common to all discovery processes e.g. methods for incorporating domain knowledge into the discovery process, dealing with missing values etc. Thirdly, the framework is inherently parallel and therefore discovery processes developed using this framework are parallel [ANAN95] and efficient for large data sets. Fourthly, it is easy to add new discovery techniques and facilities to the framework.

Data in the framework is considered to be evidence of the existence of knowledge and is represented in the form of mass functions. A mass function is defined as:

$$m{:}2^{A_1} X 2^{A_2} X \ldots X 2^{A_n} \rightarrow [0,1]$$

where the A_is are the frames of discernment of each of the attributes in the table
e.g. the tuple <Morrison, 14, India St., Belfast,NULL> is represented as

$$m<\{Morrison\},\{14\},\{India St.\},\{Belfast\},A_5> = s_i$$

where s_i is the ratio of the number of occurrences of the tuple to the total number of tuples.

A rule induced from the database is of the form A → C where A is called the Antecedent and C the Consequent of the rule. Associated with each rule there are three measures :

- *Uncertainty* : The uncertainty associated with a rule is the ratio of the number of tuples in the database that satisfy both A and C to the number of tuples in the database that satisfy only A.
- *Support* : The support for a rule is the ratio of number of tuples of the database that satisfy both the Antecedent and the Consequent to the total number of tuples in the database.
- *Interestingness* : Clearly, the number of rules that can be extracted from large 'data mines' is probably as large if not larger than the actual amount of data in the database. We, therefore, need a method for measuring the degree

of interest of a rule induced and only store a rule if its interestingness measure is greater than a threshold value. A number of indices for the interestingness of a rule have been suggested, e.g. J-measure [SMYT91] based on Information Theory and Piatetsky-Shapiro [PIAT91]. At present we use the index given by Piatetsky-Shapiro [PIAT91] defined as follows :

$$p(y)(p(x|y) - p(x))$$

where , x represents the Consequent of the rule
and y represents the Antecedent of the rule

Rules discovered from this data are represented in the form of *rule mass functions*

$$M : 2^{A} \text{ X } 2^{C} \rightarrow [0,1] \text{ X } [0,1] \text{ X } [0,1]$$

satisfying

1. $M(\langle\phi,\phi\rangle) = (0,0,0)$
2. $\sum_{X \subseteq C} M[1](\langle Y,X\rangle) = 1 \qquad \forall\, Y \subseteq A$

   $\sum_{Y \subseteq A,\, X \subseteq C} M[2](\langle Y,X\rangle) = 1$

where 2^{A} and 2^{C} are the antecedent and consequent discernment spaces
and M[1], M[2] and M[3] are the uncertainty, support and interestingness of the rule

The knowledge discovery process consists of the application of operators on mass and rule mass functions as defined above. At present five classes of operators have been identified within our framework. These are the combination, induction, domain, update and statistical operators.

Conclusions

Evidence Theory clearly shows promise in the area of uncertainty handling for Knowledge Discovery in Databases. It allows an expression for *ignorance* which is an intuitive way for dealing with missing values. Furthermore, the ability to combine different pieces of evidence at various levels of coarseness has clear advantages in discovery of knowledge from distributed, parallel and heterogeneous databases apart from allowing the development of parallel algorithms for database mining. The framework for Knowledge Discovery in Databases based on Evidence Theory also allows the incorporation of domain knowledge into the discovery process by using evidential operators. The framework is inherently parallel and discovery processes using this framework will, therefore be parallel making them efficient for large data sets.

Acknowledgements

The work presented here was funded by the IRTU, Northern Ireland. The Northern Ireland Housing Executive, ICL Ltd. and Kainos Software Ltd. were the industrial partners. The authors would also like to thank the University of Ulster Database Mining Interest Group for the numerous discussions on issues presented in this paper.

References

[ANAN94] S.S. Anand, D.A. Bell, J.G. Hughes A General Framework for Database Mining Based on Evidence Theory Internal Report, Faculty of Informatics, University of Ulster, November, 1994.

[ANAN94a] S.S. Anand, D.A. Bell, J.G. Hughes Aspects of Uncertainty Handling for Knowledge Discovery in Databases Internal Report, Faculty of Informatics, University of Ulster, November, 1994.

[ANAN95] S.S. Anand, C. M. Shapcott, D.A. Bell, J.G. Hughes Data Mining in Parallel Proc. of the WOTUG Conference, April, 1995.

[BELL93] D. A. Bell From Data Properties to Evidence, IEEE Transactions on Knowledge and Data Engineering, Vol. 5, No. 6, Special Issue on Learning and Discovery in Knowledge - Based Databases, December, 1993.

[BELL94] D.A. Bell, S.S. Anand, C.M. Shapcott Database Mining in Spatial Databases International Workshop on Spatio-Temporal Databases , 1994 .

[CODD70] E. F. Codd A relational model of data for Large Shared Data Banks CACM 13, No. 6, June, 1970.

[FAYY94] U.M. Fayyad, R. Uthurusamy AAAI-94 Workshop on Knowledge Discovery in Databases, July, 1994.

[GUAN91] J. Guan, D. Bell Evidence Theory and its Applications vol. 1 North-Holland, 1991.

[GUAN92] J. Guan, D. Bell Evidence Theory and its Applications vol. 2 North-Holland, 1992.

[MITC83] R.S. Michalski, J.G. Carbonell, T.M. Mitchell Machine Learning: An Artificial Intelligence Approach Tioga Publishing Company, Palo Alto, California 1983.

[PIAT91] G. Piatetsky-Shapiro, W.J. Frawley Knowledge Discovery in Databases, AAAI/MIT Press, 1991.

[PIAT93] G. Piatetsky-Shapiro AAAI93 Workshop on Knowledge Discovery in Databases, July, 1993.

[SHAF76] G. Shafer A Mathematical Theory of Evidence Prinston University Press, Prinston New Jersey, 1976.