DATABASE MINING IN THE NORTHERN IRELAND HOUSING EXECUTIVE

Ian C Magill, Mee G Tan, Sarabjot S Anand, David A Bell, John G Hughes

Introduction

In this paper we describe some work being carried out by the Database Mining Research Group[DMRG] within the School of Information and Software Engineering, at the University of Ulster, Jordanstown. In particular we outline how we have been investigating the use of data mining techniques on data sets from the Northern Ireland Housing Executive[NIHE].

NIHE are responsible for the management of the public sector housing stock within Northern Ireland. They are the largest such body within the European Union, and are acknowledged as leaders within their field in the use of information technology.

Within NIHE there are numerous databases in use. These fall into two main categories. There are the day to day, operational databases such as the Prawl rent accounting system and the Repairs maintenance management system. Such databases are used both for transaction processing and management information purposes. The other main category of database in use are the various data sets used by the Research Department. These are used for analysis that feeds into the formulation of policy. An example of such a data set is the five-yearly House Condition Survey. This is generated by a survey of a sample of approximately 10,000 properties across Northern Ireland. For each property up to 4,000 attributes are surveyed.

With such a large and growing amount of data within their organisation, NIHE are interested in techniques that will help them make better use of it. To this end, they have agreed to collaborate with the DMRG in testing the use of data mining techniques on their data.

It was decided in the first instance to use the House Condition Survey data set mentioned above as the target data set. Several data mining techniques being explored by the DMRG are being evaluated using this data.

The techniques in question are algorithms for the discovery of classification and association rules. We are also interested in sequential or temporal rules but our work on this is less advanced.

Ian C. Magill is employed by Kainos Software Ltd, Belfast.

Mee G. Tan, Sarabjot S. Anand and David A. Bell are with the School of Information and Software Engineering, University of Ulster at Jordanstown, Northern Ireland.

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

Classification Rules

This involves finding rules that partition data into disjoint groups. For example, consider the problem of tenant placement. We require rules involving neighbourhood and tenant attributes that predict the value of the attribute for length of tenure. If the length of tenure attribute can take the values long, average and short, we might discover the following rule.

```
if estate_type = D
and retail_facilities = nearby
and public_tranport_access = good
and no_of_children = 0
and tenant_age > 55
and tenant_car_owners = no
then length_of_tenure = long
```

Our classification algorithm is based on an existing technique [AGRA92]. We are attempting some enhancements to this in the areas of performance and handling of noisy data.

Association Rules

A more general rule relating items within a database is called an association rule. Classification rules are a particular type of association rule in which the rule has a single consequent attribute.

The following is an example of an association rule.

```
if district = BT7
and unemployed = No
and car_owner = yes
then heating_type = oil
and prop_value > 50000
and garage = yes                strength = 0.91, support = 0.01
```

Each such rule has a strength value and a support value.

- The strength value specifies the validity of the rule, ie. the proportion of cases in which the antecedent implies the consequent.

- The support value is the fraction of the database to which the rule applies.

We have been developing two distinct algorithms for discovering association rules. One approach generalises an existing algorithm[AGRA93], so that it handles multivalued attribute types. The other approach is a novel parallel technique based on the mathematical theory of evidence[ANAN95].

Objectives

The DMRG has set itself two main objectives for its work. The first objective is to devise improved data mining algorithms, in particular for mining classification and association rules. Some of the areas of improvement being investigated are :

- More efficient algorithms.

- Widening the scope of algorithms.

- Handling of noisy data.

- Incorporation of prior knowledge into algorithms.

- Measures of interestingness.

The second objective is to evaluate our algorithms on real world data sets, where real world implies:

- Very large databases.

- Databases with many tables, many attributes and many attribute types.

- Noisy data, missing and inaccurate data values.

The HCS data set fulfils the above requirements.

Mining Kernel System

To facilitate our work on data mining algorithms we have designed and built a database mining toolkit, the Mining Kernel System, or MKS.

MKS is a toolkit for implementing and evaluating database mining algorithms. It is intended primarily as a workbench for investigating algorithms, but can also be considered a research prototype for a database mining system.

MKS has two main features :

- A set of flexible components for building algorithms. Most algorithms have common components so we avoid duplicating effort.

- Flexible access to a variety of databases. This allows us to work easily with any data set from a relational database.

The virtual view mechanism of MKS allows users to create a virtual dataspace on which to apply their data mining algorithms. The space takes the form of a relation. The attributes that make up the relation can be mapped to a variety of underlying relational databases. The mapping between the attributes of a view and the attributes of the underlying data source is specified in a view definition text file. The view mechanism enables any data in a relational

database to be brought into a standard memory data structure by simply creating a view definition text file.

Initial results

The MKS toolkit has been completed and the three algorithms described above have been implemented with its facilities. Initial trials of the algorithms using the HCS data set are taking place with a formal data mining study scheduled to take place in February.

The initial trials suggest that the performance of the algorithms is reasonably linear with respect to the number of tuples processed, but much less so with respect to both the number of attributes and the number of attribute values. Also clear at this stage is the problem of interpretation of the very large number of rules that can be generated if the algorithms are not constrained to particular target rules.

Acknowledgements

We wish to thank the Northern Ireland Industrial Research and Technology Unit for providing financial support for this work; John McPeake and his staff in the NIHE Research Department for their collaboration and encouragement.

References

[AGRA92] R. Agrawal, S. Ghosh, B. Iyer, T. Imielinski, and A. Swami: An Interval Classifier for Database Mining Applications, VLDB-92, Vancouver, British Colunbia, 1992, 560-573.

[AGRA93] R. Agrawal, T. Imielinski, A. Swami: Mining association rules between rules between sets of items in large databases, Proc. of the ACM SIGMOD Conference on Management of Data, Washington D.C., May 1993.

[ANAN95] S. Anand, C. M. Shapcott, D. A. Bell, and J. G. Hughes: Data Mining in Parallel, WOTUG-18, Manchester, April 1995