Experiences of using Data Mining in a Banking Application[1]

R. I. Scott S. Svinterikou C. Tjortjis J. A. Keane

Department of Computation,
UMIST,
PO Box 88, Manchester,
M60 1QD, UK
e-mail: {rscott, sophies, christos, jak}@co.umist.ac.uk

Abstract

In recent years the ability to generate, capture and store data has increased enormously. The information contained in this data can be very important. It is recognised that, to effectively compete in increasingly competitive global markets, banks must better understand and profile their customers. An unambiguous perspective on the behaviour and attributes of customers comes from their financial history. This data can be used to enable banks to acquire and maintain good customers, where good customers are the most profitable ones.

Knowledge Discovery in Databases (KDD), often called *data mining*, is the inference of knowledge hidden within large collections of operational data. This paper reports on experiences of applying the KDD process in a banking domain. A number of data mining techniques have been used, within the KDD process, and the results obtained have influenced the business activities of the banks. The procedures used are analysed with respect to the domain knowledge they utilise, in order to evaluate the input from a domain expert during the KDD process.

1. Introduction

The amount of data collected by businesses has grown rapidly in recent years. Existing statistical data analysis techniques find it difficult to cope with the large volumes of data now available. Neither do they harness effectively the increased processing power now available. This explosive growth has lead to the need for new data analysis techniques and tools in order to find the information hidden in this data. Consequently, the research field of Knowledge Discovery in Databases (KDD) has arisen.

Banking is an area where vast amounts of data are collected. This data can be generated from bank account transactions, loan applications, loan repayments, credit card repayments, etc. It is suspected that valuable information on the financial profile of customers is hidden within these massive operational databases and that this information can be used to improve the performance of the bank.

The aim of HYPERBANK (High PERformance BANKing) [2] is to integrate business modelling, data warehousing, KDD and high performance computing to enable banks to increase profitability by improving the customer profiling process. KDD is a multi-stage, iterative process. Each stage requires the use of expert knowledge about the domain. Business models and KDD are both sources of domain knowledge and so their integration will be mutually beneficial: business models of the application will supply expert knowledge to the KDD process and useful knowledge derived by the KDD process will feed back into the business models.

[1] This work is supported by ESPRIT HPCN project no. 22693.

In this paper, we outline a KDD experiment conducted on banking data and investigate the different forms of domain knowledge used. In section 2, the KDD process is discussed and the numerous steps involved are outlined. The use of domain knowledge used within this framework is investigated in section 3 and some conclusions are made in section 4 where further work is considered.

2. The Knowledge Discovery Process

Knowledge Discovery in Databases KDD is defined as: *the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data* [1]. It combines techniques from many fields, including databases, artificial intelligence, statistics and visualisation. The process itself has many stages and is highly iterative. A widely accepted model [1] of the process has the following steps:

1. *Defining the goal of the process* – developing an understanding of the application domain and the goals of the end-user.
2. *Data selection* – selecting a data set
3. *Data preparation* – this may involve the removal of noise from the data, handling missing fields, using transformation methods to reduce the search space, deriving new attributes, etc.
4. *Choosing the data mining task* – this decision can depend upon the goal of the KDD process, the type of data available (e.g. it may be ordered) and the available techniques. Data mining tasks include discovering association rules, discovering sequential patterns, discovering similar time sequences, predicting a classification, discovering clusters and predicting values.
5. *Choosing the data mining algorithm(s)* – a data mining task may have more than one available algorithm. The choice of data mining algorithm depends on the goal of the process, i.e. whether it is predictive, descriptive, etc.
6. *Data mining* – searching for patterns of interest in the dataset.
7. *Interpreting mining results* – the presentation of the data mining results are important, as evaluation is difficult. Different visualisation techniques may be used.
8. *Consolidating discovered knowledge* – incorporating derived knowledge into the organisation.

KDD is an iterative process: the results of one step may mean that a previous step needs to be revisited.

Although the data mining step is usually the most computationally expensive, the quality of the results achieved by the process depends greatly on the other components. The choices made in these steps depend on the domain knowledge of the user and can have a large influence on the quality of the outcome of the KDD process,

3. Knowledge Discovery Experiment

The KDD experiment was performed on over 1 million records with over 50 attributes per record. Approximately 50% of the attributes were categorical with more than two possible values, 20% were binary and the remainder was continuous. The data was loaded as a single table on DB2 and the data mining tools used were IBM's Intelligent Miner and XpertRule Profiler from Attar Software. The experiment was performed with access to domain experts. The data itself was reasonably clean but some attribute values were missing and some were inconsistent. The process was analysed according to the model given in section 2.

3.1. Defining the Goal of the Process.

The goal of the process was to build a predictive and descriptive model of a customer according to some abstract measure. The first task, therefore, was to concretely define this measure and decide upon a strategy for the process.

3.2. Data Selection.

Not all of the records in the database were relevant to our experiment; the size of the

dataset was 286816 records. Initially, the number of attributes was reduced to 30 as some of the attributes were redundant and some were obviously not predictive, for example *customer_surname*.

3.3. Data Preparation.

Some attributes contained empty values. In some cases these were replaced with the default value that existed for that attribute, in other cases a default value was decided upon. The goal of the process involved the construction of a model according to some measure. There was no single attribute in the database that related directly to this measure and so a new attribute was added to the dataset that was derived from several existing attributes. More attributes were added during the experiment when it was decided that they were useful.

Some of the categorical attributes had many values. This meant that the data mining results concerning these attributes were difficult to interpret and so it was decided that the number of distinct categories be reduced by grouping attribute values. Grouping was either done automatically by the data mining tool or on the advice of domain experts.

3.4. Choosing the Data Mining Task.

The nature of the data meant that the *clustering* and *classification* techniques were used.

3.5. Choosing the Data Mining Algorithm(s).

XpertRule Profiler's decision tree induction algorithm was used. *Intelligent Miner* offers a variety of algorithms for each technique. When clustering, because of the mainly categorical nature of the data, the demographic algorithm was chosen. The goal of the KDD process was to produce a descriptive model and so the decision tree classification algorithm was chosen as opposed to a neural network based approach, which is not very descriptive.

3.6. Data Mining.

The data mining phase is the most computationally expensive and is highly iterative. Firstly, the clustering algorithm was used in order to visualise the data. Many subsets of the data were selected with the measure attribute weighted strongly, which meant that each cluster produced contained all the records for a particular measure attribute value. This gave us more of a 'feel' for the data and, to some extent, guided choices made during the clustering and classification exercises. It also highlighted some inconsistencies in the database that had to be dealt with and showed the need for grouping within an attribute.

A decision tree algorithm was used to classify the dataset with respect to the measure attribute. This process is necessarily explorative and so a trial and error approach was employed. The set of attributes used as the active data set, against which the measure attribute was classified, was constantly changing. Domain experts informed the process as to which types attributes were more 'interesting' in terms of a predictive model, i.e. information the bank could use to target customers. These attributes were deemed of greater importance than others during the iterative data mining step of the KDD process.

3.7. Interpreting Mining Results.

The graphical methods of result representation used by Intelligent Miner and XpertRule Profiler are quite easy to read. Some findings could be interpreted by a non-expert using general knowledge whereas other results had to be explained by an expert with a more intimate knowledge of the data.

4. Conclusions and Further Work.

The KDD process has many stages, as shown above. Each step has considerable influence on the overall quality of the process and requires considerable knowledge of domain. The preparation of the data set is not a trivial task. Errors in the data have to be identified, a strategy for removing these errors has to be decided upon and the data has to be cleaned according to this strategy.

The data mining process involves searching for patterns in the target data set. The

data mining algorithm can be made more useful (its efficiency can be improved and the results obtained can be simpler and easier to read) by reducing the size of this data set, whilst it is important not to remove any information held in the data. Knowledge of any relationships between data attributes can be useful when reducing the set of attributes. These relationships can be meaningful, causal or functional [3]:

- A *meaningful* relationship between attributes A and B means that A can only be understood in the context of B.
- A *causal* relationship between attributes A and B denote that A causes B.
- A *functional* relationship between A and B means that one of the two attributes can be discarded during the data preparation phase because they contain the same information.

Information on functional dependencies was used in the data selection phase to reduce the number of attributes. Knowledge of these relationships was also useful when cleaning the data as some of the attribute values involved in these functions were erroneous and it was useful to know which attribute was derived and which was recorded.

Grouping attribute values can reduce the size of the search space and can also make the data mining results much easier to read. Generalising in this way can, however, lead to loss of information. The approach taken here was, initially, to leave attribute values ungrouped and group attributes later in the experiment when required.

By analysing a simple KDD experiment in the banking domain, we have shown the extensive role that domain knowledge plays in every step of the KDD process. In this case, the information was supplied by banking domain experts. The aim of HYPERBANK is to integrate data mining, data warehousing, business modelling and high performance computing technologies to enable banks to increase profitability by the improved use of the vast amounts of customer-related data they hold.

The quality of results obtained from data mining tools depends greatly on the pre-processing performed on the data. This pre-processing relies on the effective use of domain knowledge. One of the innovations of the HYPERBANK project is the integration of domain-dependent knowledge and data mining in order to support the data preparation process.

The interpretation of data mining results is another step in the KDD process that relies heavily on domain knowledge. Often, this interpretation is based on the intuition of a domain expert and is therefore difficult to model. This process, however, can be improved if the domain expert has a thorough knowledge of the data, i.e. what all the attribute values mean, what is the source of the data, how accurate is the data, etc. Another possible approach is to use the business models to filter uninteresting results.

Business models should accurately represent the domain. The data mining results themselves are a source of domain knowledge and therefore can be used to validate and possibly enrich the business models. This cross validation highlights how the integration of KDD with business modelling will mutually benefit both areas.

Acknowledgements

The authors would like to thank all their colleagues involved in the HYPERBANK project. Thanks are also due to Attar Software and to IBM.

References

1. Fayyad, U. M., Piatetsky-Shapiro, G. and P. Smyth. “From Data Mining to Knowledge Discovery: An Overview”, in *Advances in Knowledge Discovery and Data Mining*, AAAI Press, 1996.

2. Keane, J. A. “High Performance Banking”, Proceedings of RIDE ’97, IEEE Press, 1997.

3. Cleary, J., Holmes, G., Cunningham, S. J., and Witten, I. H., “Metadata for Database Mining” Proceedings IEEE Metadata Conference, Silver Spring, MD, April 16-18, 1996.