Data Mining in Economics, Finance and Marketing

Hans C. Jessen
International Head of Econometrics
Initiative Consulting
London SW1V 1PX
UK
hans.jessen@initiativemedia.com

Georgios Paliouras
Institute of Informatics and Telecommunications
NCSR "Demokritos"
15310 Ag. Paraskevi
Athens, Greece
paliourg@iit.demokritos.gr

INTRODUCTION

Data Mining has become a buzzword in industry in recent years. It is something that everyone is talking about but few seem to understand. There are two reasons for this lack of understanding: First is the fact that Data Mining researchers have very diverse backgrounds such as machine learning, psychology and statistics. This means that the research is often based on different methodologies and communication links e.g. notation is often unique to a particular research area which hampers the exchange of ideas and the dissemination to the wider public. The second reason for the lack of understanding is that the main ideas behind Data Mining are often completely opposite to mainstream statistics and as many companies interested in Data Mining already employ statisticians, such a change of view can create opposition.

There are many definitions of Data Mining, the one we favour can be summarised as follows:[1]

> "*Data Mining is concerned with secondary data analysis of large data bases where the aim is to identify unsuspected relationships of interest or value*."

Classical statistics is mostly based on hypothesis testing. The researcher makes assumptions about the structure of the data and then uses statistical tests to either prove or disprove these assumptions. The result of such an exercise is that a lot of careful consideration goes into building a model and that the researcher should have a good understanding of the data involved. The drawback is, of course, that the quality of a model becomes dependent on the quality of the researcher, his ability to formulate interesting hypotheses and his experience in handling a given date source.

Data Mining does not have hypotheses testing at its heart and this is its main difference from classical statistics. Instead, Data Mining aims to find interesting relationships within the data that are of value to the researcher. The most appealing aspect of Data Mining is that it removes the need for the researcher to be an expert in model building and therefore reduces the cost of the analysis. It also offers the possibility that the tools might come up with ideas that the researcher would not have thought of. Although this sounds excellent for applied researchers there are unfortunately also some drawbacks. Here are a couple that are of particular interest:

Are the identified relationships of interest? Most Data Mining is used on medium or large-scale databases. With a large enough number of observations, it is all too easy to identify spurious or obvious patterns. Spurious patterns are caused by pure chance and do not relate to the general structure of the data, whereas obvious patterns are relationships in the data caused by data collection procedures or inherent in a particular type of data e.g. the colder it is, the fewer ice creams are sold. Many examples have been presented where Data Mining techniques have come up with solutions that are trivial for an expert in the field.

Selection bias. Much work has been done on selection bias in statistics but Data Mining research has largely ignored this issue. One way to think about selection bias is "how did the people on which I have data happen to be in the data set in the first place?" One often sees forecasts, say from decision trees, being applied to the whole population without recognising that the data from which the tree was derived was non-random.

Few people in industry doubt that Data Mining is here to stay and that it offers significant improvements over classical analysis when used on large databases. The challenge for the Data Mining research community is to incorporate knowledge from other fields, like econometrics and statistics, in order not to make obvious mistakes

[1] This note draws on ideas from Prof. David J. Hand's RSS (Royal Statistical Society) presentation: "Data Mining: puff or potential".

or to re-invent the wheel. The ACAI'99 Workshop on 'Data Mining in Economics, Finance and Marketing' aimed to address this challenge, by bringing together people from different disciplines who share a common interest for developing and using data mining techniques. This report presents an overview of the papers presented in the workshop and some of the general conclusions that were drawn in the course of the workshop.

OVERVIEW OF PAPERS

A. Feelders and H. Daniels: Discovery in practice

Data mining or Knowledge Discovery in Databases (KDD) is an exploratory and iterative process that can be decomposed into a number of stages. This paper describes the different activities in the data mining process and discusses some pitfalls and guidelines to circumvent them. Despite the predominant attention for analysis, data selection and pre-processing is usually the most time-consuming activity, and has a substantial influence on the ultimate success of the process. The involvement of a subject area expert, data mining expert as well as a data expert is critical to the success of data mining projects. Despite the attractive suggestion of "fully automatic" data analysis, knowledge of the processes behind the data remains indispensible in avoiding the many pitfalls of data mining. Although company databases are usually quite large, proper formulation of the data mining problem combined with sampling techniques often allows reduction to manageable sized data sets. In the majority of applications the data were originally not collected with the intention of data mining, but merely to support daily business processes. This may give rise to low quality data, as well as biases in the data that reduce the applicability of discovered patterns.

M. Dikaiakos: FIGI: Using Mobile Agent Technology to Collect Financial Information on Internet

This paper presents the architecture of a Financial Information Gathering Infrastructure (FIGI). FIGI helps investors collect, filter, combine and integrate portfolio-related information, provided through various Internet services, like World-Wide Web sites and Web-databases. FIGI is being developed with Java-based Mobile Agent technology by Mitsubishi Electric Information Technology Center [1]. The employment of Java and Mobile Agents provides a framework for unifying the various financial information services currently available on Internet and for sustaining continuous information provision to mobile users.

Melina Karamanlidou, Olivier Tuffier, Ioannis Vlahavas: Stock Miner: A System for Knowledge Discovery in Financial Data

The vast improvement of hardware and software technology in the last years has made it possible for companies to store large amounts of data in a reliable and inexpensive way. Although this may appear to be very positive at first, it has led in very many cases to large collections of data, where it has become impossible for humans to maintain the right understanding of the data they hold. This paper concentrates on data related to stock markets and especially on the stock market in Athens. It presents the system Stock Miner, which intends to deal with problems related to large volumes of data in financial areas. Stock Miner combines technical analysis, which is used by stockbrokers, and Knowledge Discovery in Databases, which is a new field concerned with analyzing data and discovering useful information.

Zdzislaw Piasta: Analyzing business databases with the ProbRough rule induction system

The approach introduced in this paper for analysing business databases is based on the idea of rule induction. The ProbRough system [5] for inducing rough classifiers was inspired by the methodology of the rough set theory. The search strategy of ProbRough, through the set of partitions of the attribute space, is guided by the global cost criterion. Because of the specific shape of partition elements, the resultant rough classifiers may be presented as sets of simple and transparent decision rules, easily understood by humans. The domains of rules are disjoint and fill up the whole attribute space. ProbRough accepts databases with noisy and inconsistent information delivered by attributes of any mixed qualitative and quantitative type. Moreover, it enables the use of background knowledge in the form of prior probabilities of decisions and different costs of misclassification. The ProbRough system is capable of inducing decision rules from databases with practically unlimited number of objects and attributes. The paper presents the behavior of the ProbRough system on several real-life business databases. Two real-world marketing databases [3,7] are being used to illustrate the way of discovering the decision rules for identifying customers who are likely to accept or reject an offer. Furthermore, the US Census

Bureau database is being analysed and applications of ProbRough in credit evaluation and financial ratio analysis are being illustrated.

Panagiotis Adamidis, Konstantinos Koukoulakis: Evolutionary Data Mining applied to TV Databases: A First Approach

The exploration of huge quantities of stored data and the extraction of useful knowledge, which is formally referred to as data mining, is now believed to be a critical factor in the decision that a company or any other interested individual may take. Recently, Evolutionary Algorithms (EAs) have been applied with very good results to various types of data mining problems. EAs are stochastic search techniques that explore combinatorial search spaces, using simulated evolution. The primary objective of an EA is to either find something – whether this is known or not – or accomplish a goal, or more generally solve a problem. This paper presents initial results of data mining from TV program databases using EAs. It compares the performance of different operators and different operator parameters of EAs. The EA mining system is used to extract rules from a database that contains TV broadcast data. The database has historical data, i.e., broadcasted schedules. Therefore, the interested individual, that is the owner of the station, given the rules, can be informed about possible relations between attributes and plan future schedules accordingly. Initial results show that EAs can discover previously unknown knowledge in the TV database that we used.

Filip Coenen, Gilbert Swinnen, Koen Vanhoof, Geert Wets: The Improvement of Response Modelling: Combining Rule-Induction and Case-Based Reasoning

Direct mail is a typical example for the use of response modelling. In order to decide which people will receive the mailing, potential customers are divided into two groups or classes (buyers and non-buyers) and a response model is created. The main aim of this paper is the improvement of response modelling. For this purpose, a combined approach of rule-induction and case-based reasoning is being proposed. The initial classification of buyers and non-buyers is done by means of the C5 algorithm, the more recent version of C4.5 [6]. In order to improve the ranking of the classified cases, a new method, called rule-predicted typicality, is being introduced. The combination of these two approaches is tested on synergy by elaborating a direct mail example.

Nikolaos Thomaidis, George Dounias, Costas D. Zopounidis: A fuzzy rule based learning method for corporate bankruptcy prediction

Corporate bankruptcy prediction is a usual problem in financing and management. Several approaches have been proposed for the solution of this problem. These can be classified into two categories: conventional classification (e.g. discriminant analysis and multicriteria analysis) and data mining methods (e.g. neural networks, genetic algorithms, decision trees). In this paper a new approach is proposed based on Machine Learning and Fuzzy Logic. This new approach uses a fuzzy system, in order to classify firms into efficient and inefficient ones. The system uses the *Fuzzy- ROSA (Rule Oriented Statistical Analysis)* [2] methodology, which is found in *Winrosa*© [4] and is a combination of the *Standard-ROSA* methodology and Fuzzy Logic. It is shown that the proposed method produces better results than the conventional ones. Furthermore, it produces rules with fewer statements in their premise, which are more 'general', and thus have greater prediction capability.

CONCLUSIONS

The workshop papers clearly showed how Data Mining can be applied to many different economic and/or financial prediction problems. What is perhaps the most interesting observation is just how quickly this research area has become popular in the otherwise conservative world of business. Statistical analysis has been available to businesses for years but somehow Data Mining has captured the interest of businesses in a way that classical statistical analysis never did. Cynics may believe that this is due to better marketing of Data Mining with labels such as "artificial Intelligence, Neural Networks, Decision rules" instead of the usual statistical labels like "nonlinear regression, discriminant analysis". This may be partly true but without a real financial benefit to businesses, Data Mining would not have gained so widespread popularity.

Although the future looks very rosy for Data Mining both as a research area and as a tool to increase profitability for businesses, there are still teething problems that need to be dealt with more rigorously. For example, several of the papers drew attention to data issues. Without a thorough understanding of how data is collected and pre-processed, it is unlikely that Data Mining can offer unbiased results. Understanding your data is not a glamorous undertaking and many businesses tend to ignore this and go straight to doing the analysis.

Perhaps there is a need for some caveats to the usual selling of Data Mining as do-it-all-without-thinking techniques. Another example raised at the workshop was the need to keep the output of the Data Mining analysis easily understandable. Most business leaders will be uncomfortable with things they do not at least have some understanding of and "black boxes" will never be acceptable.

The papers contained many examples of innovative technical developments. One of the most attractive things about this research area is the diversity of approaches to data analysis. In fact, Data Mining probably incorporates a greater variety of different techniques than any branch of numerical analysis. It is highly likely that this coming together of ideas and innovations from so many sources will continue to inspire developments in the Data Mining area and that businesses will be increasingly interested in what Data Mining can offer them.

REFERENCES

[1] Horizon Systems Laboratory. *Mobile Agent Computing. A white paper.* Mitsubishi Electric ITA., January 1998.

[2] Krone, A. and Kiendl, H.. Rule-based decision analysis with Fuzzy-ROSA method, *Proceedings of EFDAN'96*, Dortmund (Germany), 1996, 109-114.

[3] Kowalczyk, W., Piasta, Z. Rough-set inspired approach to knowledge discovery in business databases. In: X. Wu, R. Kotagiri, K. R. Korb, *Research and Development in Knowledge Discovery and Data Mining*, Proceedings of the Second Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD-98, Melburne, 15-17 April, Springer-Verlag, Berlin, Heidelberg, New York, 1998, 186-197.

[4] MIT GmbH. *WINROSA: Handbook*, Aachen, Germany, 1997(b).

[5] Piasta, Z., Lenarcik, A. Learning rough classifiers from large databases with missing values. In: L. Polkowski, A. Skowron, (eds), *Rough Sets in Knowledge Discovery*, Physica Verlag, 1998, 483-499.

[6] Quinlan, J.R., *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, 1993.

[7] Van den Poel, D., Piasta, Z. Purchase prediction in database marketing with the ProbRough system. In: L. Polkowski, A. Skowron, (eds), *Rough Sets and Current Trends in Computing*, Physica Verlag, 1998, 593-600.