Bulletin of the *Transilvania* University of Braşov • Vol. 3 (52) - 2010
Series V: Economic Sciences

DATA MINING MEETS ECONOMIC ANALYSIS: OPPORTUNITIES AND CHALLENGES

A. BAICOIANU[1] S. DUMITRESCU[1]

Abstract: *Along with the increase of economic globalization and the evolution of information technology, data mining has become an important approach for economic data analysis. As a result, there has been a critical need for automated approaches to effective and efficient usage of massive amount of economic data, in order to support both companies' and individuals' strategic planning and investment decision-making. The goal of this paper is to illustrate the impact of data mining techniques on sales, customer satisfaction and corporate profits. To this end, we present different data mining techniques and we discuss important data mining issues involved in specific economic applications. In addition, we discuss about a new method based on Boolean functions, LAD, which is successfully applied to data analysis. Finally, we highlight a number of challenges and opportunities for future research.*

Key words: *data mining, LAD, market, optimization, economic, statistics.*

1. Introduction

Data mining is a current-day term for the computer implementation of a timeless human activity. It is the process of using automated methods to uncover meaning from accumulated electronic traces of data. Companies with a strong consumer focus - retail, financial, communication, and marketing organizations use data mining. Data mining enables them to determine the relationships between "internal" factors such as price, product positioning, or staff skills, and "external" factors such as economic indicators, competition, or customer demographics.

The competitive advantages achieved by data mining include increased revenue, reduced costs, and much improved market place responsiveness and awareness. There has been a large body of research and practice focusing on exploring data mining techniques to solve economic problems.

Among of the various methods in data mining, we are considering in this paper the methodology of Logical Analysis of Data (LAD). Logical Analysis of Data is a supervised classification methodology, designed to identify patterns of findings or syndromes which predict outcomes. In these cases, the problem is to learn enough from the given data to predict whether a new 'patient' is prone to develop.

This paper provides a survey of the intersection of financial market, monetary policy and data mining, where LAD ensures collection of relevant evidence

[1] Department of Computer Science, *Transilvania* University of Braşov.

with minimal expenditure of effort, time or money.

We structured this paper in five sections, as follows. In Section 2, we describe in detail the principle of data mining and its market impact. In Section 3, we relate specifications about LAD method and its usage in financial models. We list a number of examples and results for data mining and in particular, for LAD, in Section 4. In Section 5, we discuss a number of trends and challenges for the future research in this area.

2. Data Mining for Increasing Economic Analysis Efficiency

We use symbolic models to reflect real-world events in an ever-broadening range of areas. By manipulating models, we can find more about the real world and so, we can get the profit. This is the power of knowledge discovery or data mining.

In this paper, we describe data mining in the context of economic application from technical and practical perspectives. Data mining techniques are used to uncover hidden patterns and predict future trends and behaviors in economic markets. It exploits the knowledge, that are held in the enterprise data store, by examining data that reveal patterns, which suggest better ways to produce profit, savings, higher quality products and greater customer satisfaction. The retrieval of patterns from data and the implementation of the lessons learned from the patterns are what data mining and knowledge discovery are all about. Data mining, by providing more information about the market, goes to the heart of the competitive advantage.

Thus, data mining offers three major advantages:

1. provides information about business processes, customers and market behaviors.
2. takes advantages from the data that could be available in operational data collections, data marts or data warehouses.
3. provides patterns of behavior, reflected in data that can drive the accumulation of business knowledge and the ability to foresee and shape future events.

Discovering successful patterns that are contained in data, but which are normally hidden, can be a formidable challenge. While large-scale information technology evolving separate transaction and analytical systems, data mining provides the link between those.

Data mining software analyzes relationships and patterns inside of stored transaction, data based on open-ended user queries. Several types of analytical software are available: statistical, machine learning and neural networks. Four types of relationships are sought for such analysis. They are:

1. Classes: Stored data used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be useful to increase traffic by having daily specials.
2. Clusters: Data items grouped accordingly with logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities.
3. Associations: For example, the beer-diaper-nuts.
4. Sequential patterns: Data are mined to anticipate behavior patterns or trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

Data mining consists of five major elements:

1. Extract, transform and load transaction data onto the data warehouse system.
2. Store and manage the data in a multidimensional database system.
3. Provide data access to business analysts and information technology professionals.
4. Analyze the data by application software.
5. Present the data in a useful format, such as a graph or table.

There is no lack of data in the modern enterprise. Therefore, the raw material for data mining and knowledge discovery is abundantly available. The data store contains records that have the potential to reveal patterns of dependencies that can enrich a wide variety of enterprise goals, missions and objectives. Retail sales can benefit from the examination of sales records to reveal highly profitable retail sales patterns.

Economic analysts can examine the records of financial transactions to reveal patterns of successful transactions. An engineering enterprise can search through its records surrounding the engineering process-manufacturing time; lot size, assembly parameters, and operator number to determine the combination of data conditions that relate to the quality measure of the device coming off the assembly line. Marketing analysts can look at the marketing data store to detect patterns that are associated with market growth or customer responsiveness. The data are freely available and the pay-offs are enormous: the ability to decrease inventory, increase customer buying propensity, drive product defects detection closer to the assembly line and so on one percent represents a truly staggering. The key to reaping the rewards of data mining is to have a cost-effective set of tools and body of knowledge to undertake the knowledge discovery.

2.1. Data Mining Benefits

Data mining aims to discover hidden knowledge, unknown patterns, and new rules from large databases that are potentially useful and ultimately understandable for making crucial decisions.

It applies data analysis and knowledge discovery techniques under acceptable computational efficiency limitations and produces a particular enumeration of patterns over the data. The insights obtained via a higher level of understanding of data can help iteratively improve business practice. Nowadays, data mining software vendors are integrating fundamental data mining capabilities into database engines, so that users can execute data mining tasks in parallel inside of the database, which reduces response time.

For a better understanding of how useful data mining is, we present some selective tasks solved by data mining:

1. Prediction: the task of learning a pattern, created from examples that using the developed model, to predict future values of the target variable.
2. Classification: the task of finding a function that maps records in one of the several discrete classes.
3. Detection of relations: the task of searching for the most influential attributes of a selected variable.
4. Modeling: the task of finding explicit formulae that describes dependencies between various variables.
5. Clustering: the task of identifying groups of records that are similar between themselves but different from the rest of the data.
6. Market Basket Analysis: the task of processing transactional data in order to find those groups of products that are

sold together well. One also searches for directed association rules that identify the best product, which can be offered with the selection of purchased products.

7. Deviation Detection: the task of determining the most significant changes in key measures from previous or expected values.

Other tasks could be target marketing, [18] customer relationship management, [15] cross selling, [20] market segmentation, forecasting, customer retention, improved underwriting, quality control, competitive analysis.

The main reason of the necessity of data mining strategy is the massive volume of data that require processing. Human analysts find themselves in an impossible situation when dealing with such overwhelming amounts of data, so they were obliged to access all kind of intelligent and automated data mining algorithms. [22] Other reasons for choosing this automatic analysis, instead of human analysis, are the inadequacy of human brain when searching for complex multifactor dependencies and the lack of objectiveness of human processed analysis.

A complete data mining process can substitute the work of professional statisticians that were highly trained but also highly paid, this way an analyst that is not a professional in statistics or programming will easily manage to extract knowledge from data.

Regardless of whether we are aware of it, our daily lives are influenced by data mining applications. For example, almost every financial transaction is processed by a data mining application to detect fraud. Increasingly, organizations are using data mining tools and applications together in an integrated environment for predictive analytics. Data mining tools are used to ensure flexibility and greatest accuracy possible. Essentially, data mining tools increase the effectiveness of data mining applications. Since no two organizations or data sets are alike, no single technique delivers the best results for everyone. Not only do data mining tools deliver in-depth techniques, nut data mining tools also deliver flexibility combinations of techniques to improve predictive accuracy.

Because data mining tools are so flexible, a set of data mining guidelines and a data mining methodology have been developed to guide the process. The Cross-Industry Standard Process for Data Mining [24] ensures your organization's results with data mining tools are timely and reliable. This methodology was created in conjunction with practitioners and vendors to supply data mining practitioners with checklists, guidelines, tasks, and objectives for every stage of the data mining process.

Hence, data mining provides very useful methods to realize efficient economic analysis that classical methods were not able to provide.

3. Logical Analysis of Data (LAD) in Economic Parameters

The idea of LAD was first described by P. Hammer in a lecture more than 20 years ago [10] and was subsequently expanded and developed. [9] The effectiveness of the LAD methodology has been validated by many successful applications of data analysis problems. The Logical Analysis of Data is an effective tool to develop accurate diagnostic and prognostic devices. Several detailed studies [5, 6] show that the accuracy of LAD compares with the best methods used in data analysis (SVM, random forests, artificial neural networks). LAD provides results that closely resemble and sometimes exceed the results of the most frequently used methods.

LAD, by its nature, makes it possible to examine automatically tens of thousands of possible interactions with high degrees of complexity, retaining only the most significant ones. It can be expected that the interactions revealed through LAD may stimulate research for a better understanding of the related cause-effect relationships. The value of LAD was reconfirmed and considerably strengthened in the real life medical applications. [2, 11, 16] In particular, some front of-the line medical centers are increasingly using LAD in the actual practice of medical diagnosis for a variety of syndromes and their ability to adopt improved LAD therapies involved a significant cost reduction. [19]

One of the specific features of LAD is that, in contrast to different "black box"-type methods, which provide the classification of new points without any explanations, LAD provides for each classification a justification of the reasons why LAD views an observation (e.g., patient, economical process) as being negative/positive. Each explanation of the specific reasons for which a particular patient is classified by LAD as, let us say positive, has two components:

1. the LAD classification presents the list of positive patterns displayed by the analyzed item, which is, of course, displayed by large proportions of the positive cases in the dataset, but by none of the negative ones;
2. evidence for the fact that the patient or the process in question does not satisfy the defining conditions of any of the negative patterns in the model.

The usefulness of such explanations in arriving at an economical decision is obvious to managers, customers, financial institutions, insurance companies, banking industry, and various government agencies involved in economical process.

3.1. Logical Analysis of Data - An Overview

LAD is a combinatory, optimization, and Boolean logic based methodology for analyzing archives of observations. It distinguishes from other classification methods and data mining algorithms by the fact that it generates and analyzes exhaustively a major subset of combinations of variables which can describe the positive/negative nature of observations (e.g. to describe solvent or insolvent banks, healthy or sick patients), and uses optimization techniques to extract models constructed with the help of a limited number of significant combinatorial patterns generated in this way. We shall sketch below very briefly the basic concepts of LAD, referring the reader for a more detailed description. [1, 5, 6, 11]

In LAD, as in most of the other data analysis methods, each observation is assumed to be represented by an *n*-dimensional real-valued vector. For the observations in the given dataset, beside the values of the n components of this vector, an additional binary *(0, 1)* value is also specified; this additional value is called the output or the class of the observation, with the convention that *0* is associated to negative observations, and *1* to the positive ones. The purpose of LAD is to discover a binary-valued function *f* depending on the *n* input variables, which provides discrimination between positive/negative observations, and which closely approximates the actual one. This function *f* is constructed as a weighed sum of patterns. In order to clarify how such a function *f* can be find we shall start by transforming the original dataset into one where variables can only take values *0* or *1*. We shall achieve this goal by using indicator variables, which show whether

the values of the variables, in a particular observation, are "large" or "small"; more precisely, each indicator variable shows whether the value of a numerical variable does or does not exceed a specified level, called cut-point. The selection of the cut-points is achieved by solving an associated covering problem. [7] By associating an indicator variable to each cut-point, the dataset is bynarized.

Positive/negative patterns are combinatorial rules, which impose upper and lower bounds of the values in the subset of input variables, satisfying the follows:

- Rich a sufficiently high proportion of the positive/negative observations, in the dataset, satisfying the conditions imposed by the pattern
- Rich a sufficiently high proportion of the negative/positive observations that violating at least one of the conditions imposed by the pattern.

The degree of the pattern is the number of variables, whose values are bounded in the definition of the pattern. The prevalence of a positive/negative pattern is the proportion of positive/negative observations covered by it. The homogeneity of a positive/negative pattern is the proportion of positive/negative observations among those covered by it. Patterns of low degree, high prevalence and high homogeneity have been shown to be the most effective in LAD applications. [6, 12]

The first step in applying LAD to a dataset is to generate the pandect, i.e., the collection of all patterns in a dataset. The number of patterns contained in the pandect of a dataset has such dimensions that can be exponentially large, in order of hundreds of thousands, possibly millions. Because of the enormous redundancy in this set, we shall impose a number of limitations on the set of patterns generated by restricting their degrees (to low values), their prevalence (to high values), and their homogeneity (to high values). These bounds are known as LAD control parameters. We add that the quality of patterns satisfying these conditions is usually much higher than that of patterns having high degrees, or low prevalence, or low homogeneity.

Several algorithms have been developed for the efficient generation of large subsets of the pandect corresponding to reasonable values of the control parameters. [5] The substantial redundancy among the patterns of the pandect makes necessary the extraction of (relatively small) subsets of positive/negative patterns, sufficient for classifying the observations in the dataset. Such collections of positive/negative patterns are called models. A model is supposed to include sufficiently many positive/negative patterns to guarantee that each of the positive/negative observations in the dataset is "covered" by (i.e. satisfies the conditions of) at least one of the positive/negative patterns in the model. Furthermore, good models tend to minimize the number of points in the dataset covered simultaneously by both positive/negative patterns in the model.

A LAD model can be used for classification in the following way. An observation (whether it is contained or not in the given dataset) which satisfies the conditions of some of the positive/negative patterns in the model, but which does not satisfy the conditions of any of the negative/positive patterns in the model, is classified as positive or negative. An observation satisfying both positive and negative patterns in the model is classified with the help of a discriminator that assigns specific weights to the patterns in the model. [7]

4. Examples and Results

4.1. Existing Applications of Data Mining in Finance

Financial markets are constantly generating large volume of data. Analyzing these data to reveal valuable information and support financial decision making present both great opportunities and grand challenges for data mining. Most financial data are random time series featuring noisy, nonlinear, and non-stationary behavior, thus making it difficult to model. Hundreds of new algorithms have been developed to segment, index, classify, and cluster time series. To date, data mining has become a promising solution for identifying dynamic and nonlinear relationships in financial data.

Here we discuss a few examples, drawn from media reports that describe how data mining is used today. Data mining applies to the entire customer life cycle from product conceptualization through customer acquisition, servicing, retention, and lifetime value optimization and to many business-to-business life-cycle activities as well.

In any area where we can collect data, we can use data mining to extract knowledge for competitive advantage. This clearly involves creating advantage in the three areas of enterprise excellence: products, customers, and operations. It has been applied to diverse financial areas including stock forecasting, [8, 13, 17] portfolio management and investment risk analysis, [23] prediction of bankruptcy and foreign exchange rate, detections of financial fraud, loan payment prediction, customer credit policy analysis, customer acquisition and customer targeting, profitability and risk reduction, loyalty management and cross-selling, [4] operational analysis and optimization, relationship marketing, customer attrition, churn reduction, fraud detection, [14,21] campaign management, business-to-business /channel, inventory, and supply chain management, market research, product conceptualization, product development, engineering and quality control, sales and sales management.

4.2. LAD in Finance

LAD method has been applied to financial applications such as banks' financial strength ratings [12], country risk rating [3], customer relationship management [15]. It was shown that LAD provides the most accurate rating model. The obtained ratings are successfully cross-validated, and the derived models are used to identify the financial variables most important for these applications. The studies also show that the LAD rating approach is objective, transparent, generalizable and it provides a model that is parsimonious and robust.

5. Challenges and Future Research

Despite the extensive research on applying data mining techniques to financial applications, this field is still evolving to meet the ever-increasing demand. Some challenges and emerging trends are identified for future research and practice in this field (choice of data mining methods and parameters, scalability and performance, unbalanced frequencies of financial data, text mining, mobile finance, web mining).

In order to improve the performance of data mining in financial applications, there is a trend for developing hybrid systems that integrate multiple data mining techniques. It was proved that, by decomposing a large problem into manageable parts, the system yields better performance in terms of computational efficiency, prediction accuracy, and generalization ability than a basic three-layer BP neural network.

We have discussed data mining in the context of economic analysis. Although data mining has been applied to

economic/finance for years, there are still many open issues and challenges that need to be carefully addressed in order to achieve effective financial management for both individuals and institutions. That being said, evolving data mining techniques have shown great potentials in financial applications and will continue to prosper in the new knowledge-based economy.

References

1. Alexe, G., Alexe, S. et al.: *Logical analysis of data – the vision of Peter L. Hammer*. In Annals of Mathematics and AI, Vol. 49, April 2007.
2. Alexe, G., Alexe, S. et al.: *Breast cancer prognosis by combinatorial analysis of gene expression data*. Rutgers Center for OR, 2007.
3. Alexe, S., Hammer, P. L. et al.: *A non-recursive regression model for country risk rating*. Rutgers Center for OR, 2003.
4. Berry, M., Linoff, G.: *Mastering Data Mining: The Art and Science of Customer Relationship Management*.
5. Boros, E., Ibaraki, T. et al.: *Logical analysis of binary data with missing bits*. In: AI 107, 1999, pp. 219–264.
6. Boros, E., Crama, Y. et al.: *Logical Analysis of Data: Classification with Justification*. In DIMACS Technical Report 2009-02, 2009.
7. Boros, E., Hammer P. L. et al.: *An Implementation of Logical Analysis of Data*. In IEEE on Knowledge and Data Engineering, 2000.
8. Boston, J.: *A measure of uncertainty for stock performance*. In: IEEE 1998 CIFEr, New York, 1998.
9. Crama, Y., Hammer, P. L. et al.: *Cause-effect relationships and partially defined Boolean functions*. In: Annals O R**16**, 1988, pp. 299–325.
10. Hammer, P. L.: *Partially defined Boolean functions and cause-effect relationships.* In: Int. Conf. on Multi-attribute Making Via OR-based Expert Systems, April 1986.
11. Hammer, P. L., Bonates, T.: *Logical Analysis of Data: From Combinatorial Optimization to Medical Application.* Rutgers Center for OR.
12. Hammer, P. L., Kogan, A. et al.: *Reverse-engineering banks' financial strength ratings using logical analysis of data*. Rutgers Center for OR, 2007.
13. John, G. H., Miller, P. et al.: *Stock selection using rule induction*. In: IEEE Expert, vol. 11, 1996, pp. 52–58.
14. Merix: *Fraud Detection*. Nutech Solution, Inc., Charlotte, NC.
15. Ngai, E., Li, X. et al.: *Application of data mining techniques in customer relationship management: A literature review and classification*. 2008
16. Reddy, A., Wang, H. et al.: *Logical Analysis of Data (LAD) model for the early diagnosis of acute ischemic stroke*. Rutgers Center for OR, 2008.
17. Refenes, A. N., Zapranis, A. D. et al.: *Modeling stock returns with neural networks*. In: Workshop on Neural Network Applications and Tools, London. U.K., 1993.
18. Stan Raicu, D: *A Data Mining Framework for Target Marketing.* CTI Chicago. Illinois 60604-2302.
19. Tan, J.: *Medical informatics: Concepts, Methodologies, Tools, and Applications.* USA, 2009.
20. Tang, H., Yang, Z. et al.: *Using Data Mining to Accelerate Cross-Selling*. In ISBIM '08, Business and Information Management, Wuhan, Hubei, China.
21. Weatherford, M.: *Mining for fraud*. In IEEE Intell. Syst., vol. 17, Jan./Feb. 2002, pp. 4–6.
22. Wu, X., Kumar, V. et al.: *Top 10 algorithms in data mining*. In Springer-Verlag London. 2007.
23. Xu, L., Cheung, Y.M.: *Adaptive supervised learning decision networks for traders and portfolios*. In IEEE/IAFE, New York. 1997.
24. http://www.crisp-dm.org/ CRISP-DM