Journal of Economics and International Finance Vol. 2(7), pp. 109-118, July 2010
Available online at http://www.academicjournals.org/JEIF
ISSN 2006-9812 ©2010 Academic Journals

Review

Application of data mining techniques in stock markets: A survey

Ehsan Hajizadeh*, Hamed Davari Ardakani and Jamal Shahrabi

Industrial Engineering Department, Amirkabir University of Technology, Tehran, Iran.

Accepted 5 July, 2010

One of the most important problems in modern finance is finding efficient ways to summarize and visualize the stock market data to give individuals or institutions useful information about the market behavior for investment decisions. The enormous amount of valuable data generated by the stock market has attracted researchers to explore this problem domain using different methodologies. Potential significant benefits of solving these problems motivated extensive research for years. The research in data mining has gained a high attraction due to the importance of its applications and the increasing generation information. This paper provides an overview of application of data mining techniques such as decision tree, neural network, association rules, factor analysis and etc in stock markets. Also, this paper reveals progressive applications in addition to existing gap and less considered area and determines the future works for researchers.

Key words: Stock market, data mining, decision tree, neural network, clustering, association rules, factor analysis, time series.

INTRODUCTION

In this paper, an overview of application of data mining techniques such as decision tree, neural network, association rules, factor analysis and etc in stock markets is provided.

Data mining, the science and technology of exploring data in order to discover previously unknown patterns, is a part of the overall process of knowledge discovery in databases (KDD). In today's computer-driven world, these databases contain massive quantities of information. The accessibility and abundance of this information makes data mining a matter of considerable importance and necessity.

Financial institutions such as stock markets produce huge datasets that build a foundation for approaching these enormously complex and dynamic problems with data mining tools.

Potential significant benefits of solving these problems motivated extensive research for years. The research in data mining has gained a high attraction due to the importance of its applications and the increasing generated information. Specifics of data mining in finance are coming from the need to:

1) Forecast multidimensional time series with high level of noise.
2) Accommodate specific efficiency criteria (e.g., the maximum of trading profit) in addition to prediction accuracy such as R2.
3) Make coordinated multiresolution forecast (minutes, days, weeks, months, and years).
4) Incorporate a stream of text signals as input data for forecasting models (e.g., Enron case, September 11 and others).
5) Be able to explain the forecast and the forecasting model ("black box" models have limited interest and future for significant investment decisions).
6) Be able to benefit from very subtle patterns with a short life time, and incorporate the impact of market players on market regularities (Boris and Evgenii, 2005).

A stock market or equity market, is a private or public market for the trading of company stock and derivatives of company stock at an agreed price; these are securities listed on a stock exchange as well as those only traded privately.

The expression "stock market" refers to the market that enables the trading of company stocks collective shares,

*Corresponding author. E-mail: hajizadeh.ehsan@gmail.com, ehsanhajizadeh@aut.ac.ir.

other securities, and derivatives. The stocks are listed and traded on stock exchanges which are entities a corporation or mutual organization specialized in the business of bringing buyers and sellers of stocks and securities together (http://en.wikipedia.org/wiki/Stock_ market).

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases (http://www.anderson.ucla. edu/faculty/jason.frand/teacher/technologies/palace/data mining.htm).

The remainder of this paper is organized as follows. Section 2 provides a brief review of the data mining techniques. Section 3 presents the application of them in stock markets. In section 4, progressive applications in addition to existing gap and less considered areas are explained determining the future works for researchers and at the end, the conclusion is presented.

DATA MINING TECHNIQUES

Data mining is an analytic process designed to explore data (usually large amounts of data - typically business or market related) in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data. The ultimate goal of data mining is prediction and predictive data mining is the most common type of data mining and one that has the most direct business applications. The process of data mining consists of three stages:

1) The initial exploration.
2) Model building or pattern identification with validation/verification.
3) Deployment (that is, the application of the model to new data in order to generate predictions) (Berry and Linoff, 2000).

In the following sections, some of the data mining techniques are described briefly.

Decision tree

Decision trees are powerful and popular tools for classification and prediction. In data mining, a decision tree is a predictive model which can be used to represent both classifiers and regression models. On the other hand, in operations research, decision trees refer to a hierarchical model of decisions and their consequences. Decision trees are also useful for exploring data to gaining sight into the relationships of a large number of candidate input variable to the target variable. Because decision trees combine both data exploration and modeling, they are a powerful first step in modeling process even when building the final model using some other techniques.

The decision maker employs decision trees to identify the strategy most likely to reach his goal. When a decision tree is used for classification tasks, it is more appropriately referred to as a classification tree. When it is used for regression tasks, it is called regression tree. The attractiveness of decision trees is due to the fact that, in contrast to neural networks, decision trees represent rules. Rules can readily be expressed so that humans can understand them or even directly used in a database access language like SQL so that records falling into a particular category may be retrieved.

A decision tree model consists of a set of rules for dividing a large heterogeneous population into smaller, more homogeneous groups with respect to a particular target variable. The target variable is usually categorical and the decision tree model is used either to calculate the probability that a given record belongs to each of the categories, or to classify the record by assigning it to the most likely class. Decision tree can also be used to estimate the value of continuous variable, although there are other techniques more suitable to that task (Michael and Gordon, 2004).

Neural network

Neural networks have been successfully applied in a wide range of supervised and unsupervised learning applications. Neural network methods are commonly used for data mining tasks, because they often produce comprehensible models. A neural network is a computational technique that benefits from techniques similar to ones employed in the human brain. It is designed to mimic the ability of the human brain to process data and information and comprehend patterns. It imitates the structure and operations of the three dimensional lattice of network among brain cells (nodes or neurons, and hence the term ‘‘neural’’).

Technology is inspired by the architecture of the human brain, which uses many simple processing elements operating in parallel to obtain high computation rates. Similarly, the neural network is composed of many simple processing elements or neurons operating in parallel whose functions are determined by network structure, connection strengths, and the processing performed at computing elements or nodes.

The network’s strength is in its ability to comprehend and discern subtle patterns in a large number of variables at a time without being stifled by detail. It can also carry out multiple operations simultaneously. Not only can it identify patterns in a few variables, it also can detect correlations in hundreds of variables. It is this feature of

the network that is particularly suitable in analyzing relationships among a large number of market variables. The networks can learn from experience. They can cope with ‘‘fuzzy’’ patterns — patterns that are difficult to reduce into precise rules. They can also be retrained and thus can adapt to changing market behavior. Even when a data set is noisy or has irrelevant inputs, the networks can learn important features of the data. Inputs that may appear irrelevant may in fact contain useful information. The promise of neural networks lies in their ability to learn patterns in a complex signal (Shaikh and Zahid, 2004).

Clustering

Clustering is a tool for data analysis, which solves classification problems. Its objective is to distribute cases (people, objects, events etc.) into groups, so that the degree of association can be strong between members of the same cluster and weak between members of different clusters.

In clustering, there is no preclassified data and no distinction between independent and dependent variables. Instead, clustering algorithms search for groups of records (the clusters composed of records similar to each other). The algorithms discover these similarities. This way each cluster describes, in terms of data collected, the class to which its members belong. Clustering is a discovery tool. It may reveal associations and structure in data which, though not previously evident, nevertheless are sensible and useful once found. The results of cluster analysis may contribute to the definition of a formal classification scheme, such as a taxonomy for related animals, insects or plants; suggest statistical models with which to describe populations; indicate rules for assigning new cases to classes for identification and diagnostic purposes; provide measures of definition, size and change in what previously were only broad concepts; or find exemplars to represent classes. Whatever business you are in, the chances are that sooner or later you will run into a classification problem. Cluster analysis might provide the methodology to help you solve it.

In clustering, there are no predefined classes and no examples. The records are grouped together on the basis of self-similarity. It is up to the user to determine what meaning, if any, to attach to the resulting clusters. Clustering is often done as a prelude to some other form of data mining or modeling (Michael and Gordon, 2004).

K-means clustering is an exclusive clustering algorithm. Each object is assigned to precisely one of a set of clusters. For this method of clustering we start by deciding how many clusters we would like to form from our data. We call this value k. The value of k is generally a small integer, such as 2, 3, 4 or 5, but may be larger.

There are many ways in which k clusters might potentially be formed. We can measure the quality of a set of clusters using the value of an objective function which we will take to be the sum of the squares of the distances of each point from the centroid of the cluster to which it is assigned. We would like the value of this function to be as small as possible.

We next select k points (generally corresponding to the location of k of the objects). These are treated as the centroids of k clusters, or to be more precise as the centroids of k potential clusters, which at present have no members. We can select these points in any way we wish, but the method may work better if we pick k initial points that are fairly far apart. We now assign each of the points one by one to the cluster which has the nearest centroid.

When all the objects have been assigned we will have k clusters based on the original k centroids but the ‘centroids’ will no longer be the true centroids of the clusters. Next we recalculate the centroids of the clusters, and then repeat the previous steps, assigning each object to the cluster with the nearest centroid etc (Max, 2007).

Association rules

In data mining, association rule is a popular and well researched method for discovering interesting relations between variables in large databases. Piatetsky-Shapiro describes analyzing and presenting strong rules discovered in databases using different measures of interestingness (Piatetsky-Shapiro, 1991). Based on the concept of strong rules, Agrawal et al. introduced association rules for discovering regularities between products in large scale transaction data recorded by point-of-sale (POS) systems in supermarkets (Agrawal et al., 1993). Association rule mining finds interesting associations and/or correlation relationships among large set of data items. Association rules shows attributed value conditions that occur frequently together in a given dataset. Mining association rules on large data sets has received considerable attention in recent years. Association rules are useful for determining correlations between attributes of a relation and have applications in marketing, financial, and retail sectors. Furthermore, optimized association rules are an effective way to focus on the most interesting characteristics involving certain attributes. Optimized association rules are permitted to contain uninstantiated attributes and the problem is to determine instantiations such that either the support or confidence of the rule is maximized.

For example, data are collected using bar-code scanners in supermarkets. Such ‘market basket’ data bases consist of a large number of transaction records. Each record lists all items bought by a customer on a single purchase transaction. Managers could use this data for adjusting store layouts, cross-selling, promotions, and catalog design and to identify customer segments based on buying patterns (http://www.resample.com

/xlminer/help/Assocrules/associationrules_intro.htm).

Factor analysis

Factor analysis is an essential step towards effective clustering and classification procedures. While the purpose of classification is to reduce the classification error when each pattern is assigned to an appropriate class, the main goal of factor analysis is to find and rank the important factors at hand which can represent the entire real world problem. Factor analysis originated in psychometrics, and is used in behavioral sciences, social sciences, marketing, financial market, product management, operations research, and other applied sciences that deal with large quantities of data. Factor analysis is often confused with principal components analysis. The two methods are related, but distinct, though factor analysis becomes essentially equivalent to principal components analysis if the "errors" in the factor analysis model are assumed to all have the same variance.

In decision analysis, more information will reduce decision errors no matter whether it is important or not. However, an increase in the number of factors will increase the computation time. Therefore, we have to select a sufficient number of factors which possess two properties: independence and importance. Two factors are independent when there is no correlation between them and a factor is said to be important if it has higher weight in ranking. Therefore, a dependent (highly correlated) and redundant (less important) factor can be removed without significant loss of information. In other words, the necessary condition of factor analysis is to find important and independent factors whereas the sufficient condition is that these factors are able to represent the complete information of a system which can be measured by the amount of undiscovered knowledge.

There are several developed approaches for factor analysis. Most conventional approach to factor analysis such as probability analysis, branch and bound method, sequential forward or backward method (SFS and SBS), etc. was developed in the 1970s and 1980s. As regards recent development, genetic algorithms (GAs) have been very useful in finding optimal solutions because a GA can search a large space with comparatively less computation time (Hsiao-Fan and Ching-Yi, 2004).

Time series

In statistics, signal processing, and many other fields, a time series is a sequence of data points, measured typically at successive times, spaced at (often uniform) time intervals. Time series analysis comprises methods that attempt to understand such time series, often either to understand the underlying context of the data points (where did they come from? what generated them?), or to make forecasts (predictions). Time series forecasting is the use of a model to forecast future events based on known past events: to forecast future data points before they are measured. A standard example in econometrics is the opening price of a share of stock based on its past performance. A time series model will generally reflect the fact that observations close together in time will be more closely related than observations further apart. In addition, time series models will often make use of the natural one-way ordering of time so that values in a series for a given time will be expressed as deriving in some way from past values, rather than from future values.

Data mining in finance not only follows this trend but also leads the application of relational data mining for multidimensional time series such as stock market time series. A. Cowan, a senior financial economist from US Department of the Treasury noticed that examples and arguments available in (Kovalerchuk and Vityaev, 2000) for the application of relational data mining to financial problems produce expectations of great advancements in this field in the near future for financial applications (Boris and Evgenii, 2005; Cowan, 2002).

Data mining is often concerned with what is happening over time. A key question is whether the frequency of values is constant over time.

Time series analysis requires choosing an appropriate time frame for the data; this includes not only the units of time, but also when we start counting from.

A time series chart has a wealth of information. For example, fitting a line to the data makes it possible to see and quantify long term trends. However, it does not give an idea as to whether the changes over time are expected or unexpected. For this, we need some tools from statistics (Michael and Gordon, 2004).

APPLICATION OF DATA MINING TECHNIQUES IN STOCK MARKETS

The stock market is a complex, no stationary, chaotic and non-linear dynamic system. Forecasting stock market, currency exchange rate, bank bankruptcies, understanding and managing financial risk, trading futures, credit rating, loan management, bank customer profiling, and money laundering analyses are core financial tasks for data mining (Nakhaeizadeh et al., 2002). Some of these tasks such as bank customer profiling have many similarities with data mining for customer profiling in other fields.

Stock market forecasting includes uncovering market trends, planning investment strategies, identifying the best time to purchase the stocks and what stocks to purchase. Financial institutions produce huge data sets that build a foundation for approaching these enormously complex and dynamic problems with data mining tools. Potential significant benefits of solving these problems motivated extensive research for years.

Application of decision tree in stock markets

Decision trees are excellent tools for making financial or number based decisions where a lot of complex information needs to be taken into account. They provide an effective structure in which alternative decisions and the implications of taking those decisions can be laid down and evaluated. They also help you to form an accurate, balanced picture of the risks and rewards that can result from a particular choice. In this section, we present some of the application of decision trees in stock markets.

In a stock market, how to find right stocks and right timing to buy has been of great interest to investors. To achieve this objective, Muh-Cherng et al. (2006) present a stock trading method by combining the filter rule and the decision tree technique. The filter rule, having been widely used by investors, is used to generate candidate trading points. These points are subsequently clustered and screened by the application of a decision tree algorithm. Compared to previous literature that applied such a combination technique, this research is distinct in incorporating the future information into the criteria for clustering the trading points. Taiwan and NASDAQ stock markets are used to justify the proposed method. Experimental results show that the proposed trading method outperforms both the filter rule and the previous method (Muh-Cherng et al., 2006).

Listed companies' financial distress prediction is important to both listed companies and investors. Jie and Hui (2008) present a data mining method combining attribute-oriented induction, information gain, and decision tree, which is suitable for preprocessing financial data and constructing decision tree model for financial distress prediction. On the basis of financial ratios attributes and one class attribute, adopting entropy-based discretization method, a data mining model for listed companies' financial distress prediction is designed. The empirical experiment with 35 financial ratios and 135 pairs of listed companies as initial samples got satisfying result, which testifies to the feasibility and validity of the proposed data mining method for listed companies' financial distress prediction (Jie and Hui, 2008).

Accurately, forecasting stock prices has been extensively studied. Jar-Long and Shu-Hui (2006) provided a proposal to use a two-layer bias decision tree with technical indicators to create a decision rule that makes buy or not buy recommendations in the stock market. A novel method designed for using two-layer bias decision tree to improve purchasing accuracy. Comparison with random purchases, the results indicate the system presented here not only has excellent out-of-sample forecasting performance, but also delivers a significant improvement in investment returns for all listed companies. Additionally, the proposed system has few parameter requirements, stable learning, and fast learning speed. Increasingly, the system presented here has high accuracy given large amounts of varied test data, with testing periods that experienced structural change including both bull and bear markets. Based on all of the above, they believe the proposed bias decision model is very flexible, modular and easily understandable (Jar-Long and Shu-Hui, 2006).

Chi-Lin Lu and Ta-Cheng Chen have employed decision tree-based mining techniques to explore the classification rules of information transparency levels of the listed firms in Taiwan's stock market. The main purpose of their study is to explore the hidden knowledge of information disclosure status among the listed companies in Taiwan's stock market. Moreover, the multi-learner model constructed with decision tree algorithm has been applied. The numerical results have shown that the classification accuracy has been improved by using multi-leaner model in terms of less Type I and Type II errors. In particular, the extracted rules from the data mining approach can be developed as a computer model for the prediction or classification of good/poor information disclosure potential.

By using the decision tree-based rule mining approach, the significant factors with the corresponding equality/ inequality and threshold values were decided simultaneously, so as to generate the decision rules.

Unlike many mining approaches applying neural networks related approaches in the literature, the decision tree approach is able to provide the explicit classification rules. Moreover, a multi-learner model constructed by boosting ensemble approach with decision tree algorithm has been used to enhance the accuracy rate in this work. Based on the extracted rules, a prediction model has then been built to discriminate good information disclosure data from the poor information disclosure data with great precision. Moreover, the results of the experiment have shown that the classification model obtained by the multi-learner method has higher accuracy than those by a single decision tree model. Also, the multi-learner model has less Type I and Type II errors. It indicates that the multi-learner model is appropriate to elicit and represent experts' decision rules, and thus it has provided effective decision supports for judging the information disclosure problems in Taiwan's stock market. By using the rule-based decision models, investors and the public can accurately evaluate the corporate governance status in time to earn more profits from their investment. It has a great meaning to the investors, because only prompt information can help investors in correct investment decisions (Jie and Hui, 2008).

Application of neural network in stock markets

This section provides some of the application of neural network in stock markets. It is nowadays a common notion that vast amounts of capital are traded through the stock markets all around the world. National economies are strongly linked and heavily influenced by the performance

of their stock markets. Moreover, recently the markets have become a more accessible investment tool, not only for strategic investors but for common people as well. Consequently they are not only related to macroeconomic parameters, but they influence everyday life in a more direct way. Therefore they constitute a mechanism which has important and direct social impacts. The characteristic that all stock markets have in common is the uncertainty, which is related to their short and long-term future state. This feature is undesirable for the investor but it is also unavoidable whenever the stock market is selected as the investment tool. The best that one can do is to try to reduce this uncertainty. Stock market prediction is one of the instruments in this process.

The main advantage of neural networks is that they can approximate any nonlinear function to an arbitrary degree of accuracy with a suitable number of hidden units (Hornik et al., 1989).

The development of powerful communication and trading facilities has enlarged the scope of selection for investors. David Enke and Suraphan Thawornwong introduced an information gain technique used in machine learning for data mining to evaluate the predictive relationships of numerous financial and economic variables. Neural network models for level estimation and classification are then examined for their ability to provide an effective forecast of future values. A cross-validation technique is also employed to improve the generalization ability of several models. The results show that the trading strategies guided by the classification models generate higher risk-adjusted profits than the buy-and-hold strategy, as well as those guided by the level-estimation based forecasts of the neural network and linear regression models (David and Suraphan, 2005).

Defu et al. (2007) dealt with the application of a well-known neural network technique, multilayer back-propagation (BP) neural network, in financial data mining. A modified neural network forecasting model is presented, and an intelligent mining system is developed. The system can forecast the buying and selling signs according to the prediction of future trends to stock market, and provide decision-making for stock investors. The simulation result of seven years to Shanghai composite index shows that the return achieved by this mining system is about three times as large as that achieved by the buy-and-hold strategy, so it is advantageous to apply neural networks to forecast financial time series, so that the different investors could benefit from it (Defu et al., 2004).

Accurate volatility forecasting is the core task in the risk management in which various portfolios' pricing, hedging, and option strategies are exercised. Tae (2007) proposes hybrid models with neural network and time series models for forecasting the volatility of stock price index in two view points: deviation and direction. It demonstrates the utility of the hybrid model for volatility forecasting. This model demonstrates the utility of the neural network forecasting combined with time series analysis for the financial goods (Tae, 2007).

Application of clustering in stock markets

As part of a stock market analysis and prediction system consisting of an expert system and clustering of stock prices, data is needed. Stock markets are recently triggering a growing interest in the physicists' community.

Basaltoa et al. (2005) apply a pair wise clustering approach to the analysis of the Dow Jones index companies, in order to identify similar temporal behavior of the traded stock prices. The objective of this attention is to understand the underlying dynamics which rules the companies' stock prices. In particular, it would be useful to find, inside a given stock market index, groups of companies sharing a similar temporal behavior. To this purpose, a clustering approach to the problem may represent a good strategy. To this end, the chaotic map clustering algorithm is used, where a map is associated to each company and the correlation coefficients of the financial time series to the coupling strengths between maps. The simulation of a chaotic map dynamics gives rise to a natural partition of the data, as companies belonging to the same industrial branch are often grouped together. The identification of clusters of companies of a given stock market index can be exploited in the portfolio optimization strategies (Basaltoa et al., 2005).

Graph representation of the stock market data and interpretation of the properties of this graph gives a new insight into the internal structure of the stock market.

Vladimar et al. (2006) study different characteristics of the market graph and their evolution over time and came to several interesting conclusions based on their analysis. It turns out that the power-law structure of the market graph is quite stable over the considered time intervals; therefore one can say that the concept of self-organized networks, which was mentioned above, is applicable in finance, and in this sense the stock market can be considered as a "self-organized" system. Another important result is the fact that the edge density of the market graph, as well as the maximum clique size, steadily increases during the last several years, which supports the well-known idea about the globalization of economy which has been widely discussed recently. They also indicate the natural way of dividing the set of financial instruments into groups of similar objects (clustering) by computing a clique partition of the market graph. This methodology can be extended by considering quasi-cliques in the partition, which may reduce the number of obtained clusters (Vladimir et al., 2006).

Stock prices tend to cluster at round numbers, a phenomenon observed in many markets. Using tick-by-tick transaction data, Wataru (2006) studies price

clustering on the Tokyo stock exchange, which is a computerized limit order market. As for the intraday pattern, the degree of price clustering is greatest at the market opening. Then, it decreases during the first half hour and reaches a stable level. It does not increase again near the market closing. There is no clear difference in clustering between call auctions and continuous auctions (Wataru, 2006).

S´onia et al. (2008) have investigated the long memory and volatility clustering for the S&P 500, NASDAQ 100 and Stoxx 50 indexes in order to compare the US and European Markets. Additionally, they have compared the results from conditionally heteroscedastic models with those from the entropy measures. In the latter, they have examined Shannon entropy, Renyi entropy and Tsallis entropy. The results have corroborated the previous evidence of nonlinear dynamics in the time series considered.

The main goal of their study is to compare two different perspectives: the so-called traditional approach in which the authors have considered the GARCH (1,1), IGARCH (1,1) and FIGARCH (1,d,1) specifications and the econophysics approach based on the concept of entropy. For their purpose three variants of this notion were chosen: the Shannon, Renyi and Tsallis measures. The results from both perspectives have shown nonlinear dynamics in the volatility of S&P 500, NASDAQ 100 and Stoxx 50 indexes and must be understood in complementarity. They have considered that the concept of entropy can be of great help when analyzing stock market returns since it can capture the uncertainty and disorder of the time series without imposing any constraints on the theoretical probability distribution. By contrast, the ARCH/GARCH type models assume that all variables are independent and identically distributed (i.i.d) (S´onia et al., 2008).

Application of association rules in stock markets

As stated in Agrawal et al. (1993) discovering association rules is an important data mining problem, and there has been considerable research on using association rules in the field of data mining problems. The associations' rules algorithm is used mainly to determine the relationships between items or features that occur synchronously in the database. For instance, if people who buy item X also buy item Y, there is a relationship between item X and item Y, and this information is useful for decision makers. Therefore, the main purpose of implementing the association rules algorithm is to find synchronous relationships by analyzing the random data and to use these relationships as a reference during decision-making (Agrawal et al., 1993).

One of the most important problems in modern finance is finding efficient ways to summarize and visualize the stock market data to give individuals or institutions useful information about the market behavior for investment decisions. The enormous amount of valuable data generated by the stock market has attracted researchers to explore this problem domain using different methodologies.

Shu-Hsien et al. (2008) investigated stock market investment issues on Taiwan stock market using a two-stage data mining approach. The first stage apriori algorithm is a methodology of association rules, which is implemented to mine knowledge and illustrate knowledge patterns and rules in order to propose stock category association and possible stock category investment collections. Then the K-means algorithm is a methodology of cluster analysis implemented to explore the stock cluster in order to mine stock category clusters for investment information. By doing so, they propose several possible Taiwan stock market portfolio alternatives under different circumstances (Shu-Hsien et al., 2008).

Application of factor analysis in stock markets

Factor analysis is particularly useful in situations where a large number of variables are believed to be determined by a relatively few common causes of variation. Also, it should be particularly useful for analyzing financial markets because if financial markets are efficient, nominal returns will be affected by default and market risk and by expected inflation and inflation uncertainty.

Michael Flad et al. provided an empirical analysis of the common factor driving the intraday movements of the DAX and the DJIA during overlapping trading hours. Based on a minute-by-minute dataset spanning from March to December, 2003, they estimate a bivariate common factor model for the two indices. By explicitly modeling the two stock indices, they implicitly assume that news on economic fundamentals is aggregated in both equity market and that, therefore, both stock indices are linked by a common trend of cumulated random information arrivals. They compute various measures of information leadership and find that the DJIA is the predominant source of price relevant information flowing into the transatlantic system of stock indices. Moreover, our impulse response analyses show that both stock markets adjust very quickly (in less than 5 min) to shocks emanating from either the U.S. or the German side, but that DJIA-innovations have a major long-run effect on the German stock market, whereas DAX-shocks are of minor importance. This observation is further strengthened by our permanent-transitory decomposition. Hence, their analysis implies that international economic news is first incorporated in the U.S. and then transferred to the German stock market (Michael and Robert, 2007).

A large and growing body of empirical work is devoted to estimating the relation between risk and return in the U.S. stock market. Although theory typically predicts a positive relation, empirical findings are mixed and often

suggest a negative relation. An important limitation of existing empirical work, however, pertains to the relatively small amount of conditioning information used to estimate the conditional mean and conditional volatility of excess stock market returns. In turn, the use of such sparse information sets in the construction of fitted moments that can translate into an omitted information bias in the estimated risk–return relation.

Sydney and Serena (2007) considered one approach to this omitted-information problem by employing a methodology for incorporating a large amount of conditioning information in their estimates of the conditional mean and conditional volatility of excess stock market returns. Recent research on dynamic factor models finds that the information in a large number of economic time series can be effectively summarized by a relatively small number of estimated factors, affording the opportunity to exploit a rich base of information more likely to span the information sets of financial market participants than in previous analyses. In doing so, their study contributes to the empirical literature by evaluating both the potential role of omitted information in the estimated risk–return relation as well as the robustness of previous results to conditioning on richer information sets (Sydney and Serena, 2007).

Application of time series in stock market

It is obvious that forecasting activities play an important role in our daily life. The traditional statistical approaches for time series can predict problems arising from new trends, but fail to forecast the data with linguistic facts. Furthermore, the traditional time series requires more historical data along with some assumptions like normality postulates.

In recent years, many researchers have used fuzzy time series to handle forecasting problems. A number of researchers presented fuzzy time series forecasting models in the last 15 years (Tahseen and Syed, 2008).

Lee et al. (2006) presented two factor high-order fuzzy time series for forecasting daily temperature in Taipei and TAIFEX. Jilani and Burney (2007a, b) and Jilani et al. (2007) presented new fuzzy metrics for high-order multivariate fuzzy time series forecasting for car road accident casualties in Belgium.

Yu proposed an appropriate approach to define the length of intervals during the formulation of fuzzy relationships and applied on TAIEX and enrollments forecasting problems (Yu, 2005a). Huarng and Yu (2005) proposed a Type-2 fuzzy time series model and applied to TAIEX forecasting problem. But their method requires extra observations to form type-2 fuzzy relations for each observation and thus require larger number of data than the conventional type-I methods (Huarng and Yu, 2005). Yu proposes weighted models to tackle fuzzy time series forecasting and applied them on TAIEX forecasting (Yu, 2005b).

Finally, Tahseen and Syed (2008) have proposed a new method for time series forecasting having simple computational algorithm. The proposed method sub-partitioned the universe of discourse based on frequency density approach using initial equal length intervals. A trend parameter was introduced that predicts the direction of the data for next observation using last three values. The trend predictor has used to adjust the weights of the proposed fuzzy metric for forecasting. The suitability of the method is examined for forecasting TAIEX and enrollments of the University of Alabama. It is clear from the results that the proposed method gives the best accuracy as compared to other methods for forecasting models (Tahseen and Syed, 2008).

Time series data is characterized as large in data size, high dimensionality and update continuously. Moreover, the time series data is always considered as a whole instead of individual numerical fields. Indeed, a large set of time series data is from stock market. Moreover, dimensionality reduction is an essential step before many time series analysis and mining tasks. For these reasons, research is prompted to augment existing technologies and build new representation to manage financial time series data. Tak-chung et al. (2008) have represented financial time series according to the importance of the data points. With the concept of data point importance, a tree data structure, which supports incremental updating, has been proposed to represent the time series and an access method for retrieving the time series data point from the tree, which is according to their order of importance, has been introduced. This technique is capable of presenting the time series in different levels of detail and facilitates multi-resolution dimensionality reduction of the time series data.

The authors have proposed different data point importance evaluation methods, a new updating method and two dimensionality reduction approaches and evaluated them by a series of experiments. Finally, the application of the proposed representation on mobile environment has been demonstrated (Tak-chung et al., 2008).

CONCLUSION AND FUTURE WORKS

With the increase of economic globalization and evolution of information technology, financial data are being generated and accumulated at an unprecedented pace. As a result, there has been a critical need for automated approaches to effective and efficient utilization of massive amount of financial data to support companies and individuals in strategic planning and investment decision making. Data mining techniques have been used to uncover hidden patterns and predict future trends and behaviors in financial markets. The competitive advantages achieved by data mining include increased revenue, reduced cost, and much improved marketplace responsiveness and awareness. There has been a large

body of research and practice focusing on exploring data mining techniques to solve financial problems.

There are many contributions on this field. Study of implementing data mining approaches and integrating them into stock market research on Tehran stock market is an example for future research and implementation. Another future extension of this research involves incorporating some other variables into the criteria for data mining techniques. These include new variables that reflect future information and those that reflect the impacts of other stock markets to the market of concern.

The research reviewed in this paper has mainly concentrated on applications of the algorithms. The quality of the data and data preparation issues, particularly relating to financial databases has not been discussed. Major effort is needed in the data preparation process, as this is often simply based on practitioner's instinct and experience. A more generic process for data cleaning is essential to enable the growth of data mining in financial market.

The stock market data mining research often does not consider the quality of the rules or knowledge discovered. The knowledge generated is sometimes cumbersome and the relationships obtained are too complex to understand. Future research effort is therefore also needed to enhance the expressiveness of the knowledge. Further research is needed to develop generic guidelines for a variety of different data and types of problems, which are commonly faced by financial markets.

To be successful, a data mining project should be driven by the application needs and results should be tested quickly. Financial applications provide a unique environment where efficiency of the methods can be tested instantly, not only by using traditional training and testing data but making real stock forecast and testing it the same day. This process can be repeated daily for several months collecting quality estimates. This paper states problems of data mining in finance (stock market) and specific requirements for data mining methods including in making interpretations, incorporating relations and probabilistic learning.

The data mining techniques outlined in this paper advances pattern discovery methods that deals with complex numeric and non-numeric data, involving structured objects, text and data in a variety of discrete and continuous scales (nominal, order, absolute and so on). Also, this paper shows benefits of using such techniques for stock market forecast.

Currently the success of data mining exercises has been reported in literature extensively. Typically it is done by comparing simulated trading and forecasting results with results of other methods and real gain/loss and stock.

We expect that in the coming year's data mining in finance will be shaped as a distinct field that blends knowledge from finance and data mining, similar to what we see now in bioinformatics where integration of field specifics and data mining is close to maturity. We also expect that blending with ideas from the theory of dynamic systems, chaos theory, and physics of finance will deepen.

REFERENCES

Agrawal R, Imilienski T, Swami A (1993). Mining association rules between sets of items in large databases, In Proceedings of the ACM SIGMOD international conference on management of data.

Basaltoa N, Bellottib R, De Carlob F, Facchib P, Pascazio S (2005). Clustering stock market companies via chaotic map synchronization, Physica A.

Berry MJA, Linoff GS (2000). Mastering data mining, New York: Wiley.

Boris K, Evgenii V (2005). Data Mining for Financial Applications, the Data Mining and Knowledge Discovery Handbook.

Chi-Lin L, Ta-Cheng C (2009). A study of applying data mining approach to the information disclosure for Taiwan's stock market investors, Expert Systems with Applications,.

Cowan A (2002). Book review: Data Mining in Finance, Int. J. forecasting.

David E, Suraphan T (2005). The use of data mining and neural networks for forecasting stock market returns, Expert Systems with Applications.

Defu Z, Qingshan J, Xin L (2004). Application of Neural Networks in Financial Data Mining, Proceedings of world academy of science, Eng. Technol.

Hornik K (1989). Stinchcombe M. and White H., "Multilayer feed forward networks are universal approximators", Neural Networks.

Hsiao-Fan W, Ching-Yi K (2004). Factor Analysis in Data Mining, Computers and Mathematics with Applications.

http://en.wikipedia.org/wiki/Stock_market

http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm

http://www.resample.com/xlminer/help/Assocrules/associationrules_intro.htm.

Huarng K, Yu HK (2005). A type 2 fuzzy time series model for stock index forecasting, Physica.

Jar-Long W, Shu-Hui C (2006). Stock market trading rule discovery using two-layer bias decision tree, Expert Systems with Applications.

Jie S, Hui L (2008). Data mining method for listed companies' financial distress prediction, Knowledge-Based Systems.

Jilani TA, Burney SMA (2007a). Multivariate stochastic fuzzy forecasting models, Expert Systems With Applications.

Jilani TA, Burney SMA (2007b). M-factor high order fuzzy time series forecasting for road accident data, in: Analysis and Design of Intelligent Systems Using Soft Computing Techniques, in: Advances in Soft Computing.

Jilani TA, Burney SMA, Ardil C (2007). Multivariate high order fuzzy time series forecasting for car road accidents, Int. J. Comp. Intelligence.

Kovalerchuk B, Vityaev E (2000). Data Mining in Finance: Advances in Relational and Hybrid Methods, Kluwer.

Lee LW, Wang LW, Chen SM, Leu YH (2006). Handling forecasting problems based on two-factor high-order time series, IEEE Transactions on Fuzzy Systems.

Max B (2007). Principles of Data Mining, Springer.

Michael JAB, Gordon S (2004). Linoff, Data Mining Techniques, for Marketing Sales, and Customer Relationship Management, WILEY.

Michael JA (2004). Berry and Gordon S. Linoff, Data Mining Techniques for Marketing, Sales, and Customer Relationship Managemanet, 2nd edn, Wiley.

Michael F, Robert CJ (2007). A common factor analysis for the US and the German stock markets during overlapping trading hours, Int. Fin. Markets, Inst. Money.

Muh-Cherng W, Sheng-Yu L, Chia-Hsin L (2006). An effective application of decision tree to stock trading, Expert Systems with Applications.

Nakhaeizadeh G, Steurer E, Bartmae K (2002). Hand book of data

mining and knowledge discovery, Oxford Univ. Press, Oxford.
Piatetsky S (1991). Discovery, analysis and presentation of strong rules, Knowledge Discovery in Databases, MIT Press, Cambridge.
Shaikh AH, Zahid I (2004). Using neural networks for forecasting volatility of S&P Shu-Hsien L, Hsu-hui H, Hui-wen L (2008). Mining stock category association and cluster on Taiwan stock market, Expert Systems with Applications. 500 Index futures prices, J. Bus. Res.
S´onia RB, Rui M, Diana AM (2008). Long memory and volatility clustering: Is the empirical evidence consistent across stock markets? Physica A 387 3826–3830.
Sydney CL, Serena N (2007). The empirical risk–return relation: A factor analysis approach, J. Fin. Econ.
Tae HR (2007). Forecasting the volatility of stock price index, Expert Systems with Applications.
Tahseen AJ, Syed MAB (2008). A refined fuzzy time series model for stock market forecasting, PHYSICA A, Physica.
Tak-chung F, Fu-lai C, Robert L, Chak-man N (2008). Representing financial time series based on data point importance, Engineering Applications of Artificial Intelligence.
Vladimir B, Sergiy B, Panos MP (2006). Mining market data: A network approach, Comput. Oper. Res.
Wataru O (2006). An analysis of intraday patterns in price clustering on the Tokyo Stock Exchange, J. Bank. Fin.
Yu HK (2005a). A refined fuzzy time-series model for forecasting, Physica.
Yu HK (2005b). Weighted fuzzy time series models for TAIEX forecasting, Physica,.