Discovering Golden Nuggets: Data Mining in Financial Application

Dongsong Zhang and Lina Zhou

***Abstract*—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. In this paper, we describe data mining in the context of financial application from both technical and application perspectives. In addition, we compare different data mining techniques and discuss important data mining issues involved in specific financial applications. Finally, we highlight a number of challenges and trends for future research in this area.**

***Index Terms*—Data mining, financial application, genetic algorithm, neural networks, rule induction, statistical inference.**

I. Introduction

With the increase of economic globalization and evolution of information technology, financial data are being generated and accumulated at an unprecedented rate. It is used to keep track of companies' business performance, monitor market changes, and support financial decision-making. Nonetheless, the rapidly growing volume of data has far exceeded our ability to analyze them manually. There is a critical need for automated approaches to effective and efficient utilization of massive financial data to support companies and individuals in strategic planning and investment decision-making.

Data mining is able to uncover hidden patterns and predict future trends and behaviors in financial markets. It creates opportunities for companies to make proactive and knowledge-driven decisions in order to gain a competitive advantage. Data mining has been applied to a number of financial applications, including development of trading models, investment selection, loan assessment, portfolio optimization, fraud detection, bankruptcy prediction, real-estate assessment, and so on. The competitive advantages achieved by data mining include increased revenue, reduced cost, and much improved marketplace responsiveness and awareness.

This paper focuses on existing data mining applications in finance. The rest of the paper is organized as follows. Section II introduces the basic concept of data mining and issues involved in data mining in financial applications. Section III describes several commonly used data mining techniques and compares their features across multiple dimensions. The data mining techniques in the context of specific financial applications are discussed in details in Section IV. In Section V, we discuss a number of trends and challenges for the future research in this area.

II. Classification and Issues of Data Mining in Financial Application

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 [1]. 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 the database, which reduces response time.

Based on the type of knowledge that is mined, data mining can be mainly classified into the following categories [2].

1) *Association rule mining* uncovers interesting correlation patterns among a large set of data items by showing attribute-value conditions that occur together frequently. A typical example is market basket analysis, which analyzes purchasing habits of customers by finding associations between different items in customers' "shopping baskets."
2) *Classification and prediction* is the process of identifying a set of common features and models that describe and distinguish data classes or concepts. The models are used to predict the class of objects whose class label is unknown. A bank, for example, may classify a loan application as either a fraud or a potential business using models based on characteristics of the applicant. A large number of classification models have been developed for predicting future trends of stock market indices and foreign exchange rates.

Manuscript received April 28, 2003; revised October 6, 2003 and December 23, 2003. This paper was recommended by Associate Editor S. Lakshmivarahan.

The authors are with the Department of Information Systems, University of Maryland, Baltimore County, Baltimore, MD 21250 USA (e-mail: zhangd@umbc.edu; zhoul@umbc.edu).

Digital Object Identifier 10.1109/TSMCC.2004.829279

3) *Clustering analysis* segments a large set of data into subsets or clusters. Each cluster is a collection of data objects that are similar to one another within the same cluster but dissimilar to objects in other clusters. In other words, objects are clustered based on the principle of maximizing the intra-class similarity while minimizing the inter-class similarity. For example, clustering techniques can be used to identify stable dependencies for risk management and investment management.
4) *Sequential pattern and time-series mining* looks for patterns where one event (or value) leads to another later event (or value). One example is that after the inflation rate increases, the stock market is likely to go down.

The knowledge to be mined is closely related to a target application and the original data. Therefore, data mining should be considered along with several other issues rather than an isolated task. First, data mining needs to take ultimate applications into account. For example, credit card fraud detection and stock market prediction may require different data mining techniques. Second, data mining is dependent upon the features of data. For example, if the data are of time series, data mining techniques should reflect the features of time sequence. Third, data mining should take advantage of domain models. In finance, there are many well-developed models that provide insight into attributes that are important for specific applications. Many applications combine data mining techniques with various finance and accounting models (e.g., capital asset pricing model and the Kareken–Wallace model). The fact that data mining in finance is involved with applications, data, and domain models leads to a conceptual framework consisting of three-dimensions, as shown in Fig. 1.

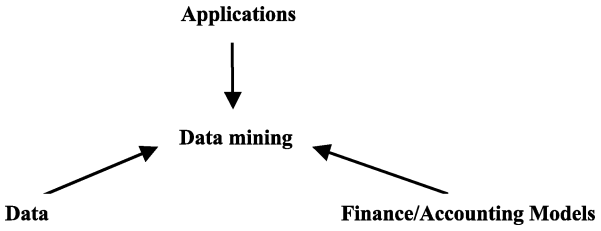


Fig. 1. Three dimensions of data mining in financial application.

III. Existing Data Mining Techniques

Among a variety of data mining techniques that have been used in finance, we mainly focus on introducing five commonly used techniques, namely neural networks, genetic algorithms, statistical inference, rule induction, and data visualization.

A. Overview of Data Mining Techniques

Neural Networks: Artificial neural networks are computer models built to emulate the human pattern recognition function through a similar parallel processing structure of multiple inputs. A neural network consists of a set of fundamental processing elements (also called neurons) that are distributed in a few hierarchical layers. Most neural networks contain three types of layers: input, hidden, and output. After each neuron in a hidden layer receives the inputs from all of the neurons in a layer ahead of it (typically an input layer), the values are added through applied weights and converted to an output value by an activation function (e.g., the Sigmoid function). Then, the output is passed to all of the neurons in the next layer, providing a feedforward path to the output layer. The weights between two neurons in two adjacent layers are adjusted through an iterative training process while training samples are presented to the network. They are used to store captured knowledge and make it available for future use. Characterized by the pattern of connections between neurons, the method of determining weights on the connections, and the node activation function, a neural network is designed to capture causal relationships between dependent and independent variables in a given data set. Neural networks offer a class of tools that can approximate financial patterns to a satisfactory degree of accuracy.

Genetic Algorithms: The basic idea of genetic algorithms is that given a problem, the genetic pool of a specific population potentially contains the solution, or a better solution. Based on genetic and evolutionary principles, the genetic algorithm repeatedly modifies a population of artificial structures through the application of initialization, selection, crossover, and mutation operators in order to obtain an evolved solution. It starts with a set of solutions (represented by chromosomes) called population. Solutions from one population are taken and used to form a new population. This is motivated by a hope that the new population will be better than the old one. Solutions used to form new solutions (offsprings) are selected according to their fitness—the more suitable they are, the more chances they have in reproduction. This evolving process is repeated many times until certain condition (e.g., the number of populations or improvement of the best solution) is satisfied.

Statistical Inference: Statistics provides a solid theoretical foundation for the problem of data analysis. Through hypothesis validation and/or exploratory data analysis, statistical techniques give asymptotic results that can be used to describe the likelihood in large samples. The basic statistical exploratory methods include such techniques as examining distribution of variables, reviewing large correlation matrices for coefficients that meet certain thresholds, and examining multidimensional frequency tables. Multivariate exploratory techniques designed specifically for identifying patterns in multivariate data sets include cluster analysis, factor analysis, discriminant function analysis, multidimensional scaling, log-linear analysis, canonical correlation, stepwise linear and nonlinear regression, time-series analysis, and classification trees. Among them, discriminant analysis, factor analysis, principle component analysis, and regression models have been frequently used to identify either influential variables in financial problems or relationships between different variables and financial markets.

Probabilistic methods assume that the construction of a model and calculations are carried out in correspondence with the probability theory. They allow users to get a probability distribution of results on the basis of variable distribution. Therefore, probabilistic methods such as the hidden Markov model (HMM) have been used in financial applications such as investment risk analysis.

Rule Induction: Rule induction models belong to the logical, pattern distillation based approaches of data mining. Based on data sets, these techniques produce a set of if-then rules to represent significant patterns and create prediction models. Such models are fully transparent and provide complete explanations of their predictions.

One commonly used and well-known type of rule induction is the family of algorithms that produce decision trees. A decision tree, which is usually constructed using a training data set, consists of hierarchically organized sets of rules. It is a simple recursive structure for representing a decision procedure in which a new instance is classified into one of the predefined classes. In decision trees, instances are represented as feature vectors containing a list of attribute-value pairs. Each internal node represents a decision attribute-value test. Each branch represents an outcome of the test, and each leaf node denotes a decision class. A credit card company, for instance, may have customer records consisting of descriptors or attributes. With a known credit history, the records may be labeled/classified as good, medium, or poor. A rule induction technique may generate a symbolic classification model that has a rule stating "If a credit card applicant has annual income higher than $40K and is between 35–55 years old and is married, then the card should be issued."

The decision tree technique is based on a divide-and-conquer approach to the classification problem. It works in a top-down manner: at each stage, it seeks an attribute that separates classes the best to split on, and then recursively processes the partitions resulted from the split. The basic principle is to maximize the entropy of the split subsets, while recursive partitioning is designed to minimize the expected cost of misclassification.

Data Visualization—"Seeing" the Data: Data are difficult to interpret due to its overwhelming size and complexity. In order to achieve effective data mining, it is important to include people in the data exploration process and combine the flexibility, creativity, and general knowledge of people with the enormous storage capacity and computational power of today's computers. Data visualization is the process of analyzing and converting data into graphics, thus taking advantage of human visual systems. This technique allows decision makers and analysts to gain insight into the data, draw conclusions, and directly interact with the data. It is proven to be of high value in exploratory data analysis, especially useful when little is known about the data and exploration goals are vague. In addition, visualization techniques can also guide researchers' intuition and provide much more intuitive ways for them to understand results. They maintain a global view of large amount of data while still preserving the perception of small regions of interest.

Typical financial applications of data visualization techniques include retail banking (e.g., product cross-selling analysis, credit risk, and electronic banking management), economic analysis, fraud detection, and portfolio performance analysis and optimization. Mutual fund companies, for example, often generate a correlation matrix. If the data set with 30 variables has 30 rows and 30 columns, the correlation matrix will include 900 entries, which is too large to view and interpret at once. A correlation image can accommodate a large number of variables while still presenting useful information [3].

TABLE I
COMPARISON OF FIVE DATA MINING TECHNIQUES

	Neural network	Genetic algorithm	Statistical inference	Rule induction	Data visualization
Ease of encoding	Low	Very low	High	Very high	Medium
Flexibility	High	Medium	Medium	Low	Low
Autonomy	High	High	Low	Low	Very high
Computation complexity	Very high	Very high	Medium	Low	Very high
Interpretability	Very low	High	Medium	Very high	Very high
Optimization capability	Medium	High	Medium	Medium	Very low
Scalability	Very high	Medium	Medium	Very low	Low
Accessibility	High	Low	Very high	High	Low

B. Comparison of Data Mining Techniques

Each data mining technique has its inherent limitations and underlying assumptions that make it a better choice for some applications but not others. We compare the above five data mining techniques on a 5-point scale ranging from very low to very high based on eight criteria: ease of problem encoding, flexibility, autonomy, computation complexity, interpretability, optimization capability, scalability, and accessibility (see Table I). Ease of problem encoding refers to the complexity in encoding a problem. Flexibility mainly concerns the ability to handle various data types and a wide range of problems. Autonomy indicates independence of prior assumptions of functional relationships between variables and of domain expertise. Computation complexity pertains to the computational cost involved in generating results. Interpretability refers to the ability to explain data mining results clearly. Optimization capability concerns generating optimal results rather than converging prematurely to an inferior solution. Scalability implies the degree of extra effort required by a data mining technique to obtain results from a larger-scaled data set. Accessibility refers to the availability of off-the-shelf software.

Since data mining is data-oriented without a strong theoretical background, data mining models are very sensitive to changes in the data and need continuous remodeling as the data or situation changes. So far, neural network modeling has been the most commonly used data mining technique in financial applications.

IV. 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 nonstationary behavior, thus making it difficult to model. A time series is a sequence of real numbers that represent values of a real variable measured at equal time intervals. For example, a time series can represent movements of stock prices or exchange rates. The conventional statistical analysis and tests indicate that financial time series has nonrandom behavior [4]. It results in the widespread use of neural networks for financial time-series prediction due to their capability of decoding nonlinear time-series

data. It is suggested that while selecting data for mining, long training durations and large samples are preferred for discovery of robust models. The time-series recency effect states that constructing models with data that are closer in time to the data that are to be forecast by the model produces higher quality. In the past decade, there has been extensive effort in mining time-series data. 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. It has been applied to diverse financial areas including stock forecasting, portfolio management and investment risk analysis, prediction of bankruptcy and foreign exchange rate, detections of financial fraud, loan payment prediction, customer credit policy analysis, and so on. In this paper, we primarily focus on the first five applications in the above list, which have mostly been discussed in the literature.

A. Prediction of the Stock Market

Investors in the market want to maximize their returns by buying or selling their investments at an appropriate time. Since stock market data are highly time-variant and are normally in a nonlinear pattern, predicting the future trend (i.e., rise, decrease, or remain steady) of a stock is a challenging problem.

Prior research has demonstrated that the prediction of future returns of individual stocks can be based on the growth rates of a number of fundamental factors such as revenues, earnings per share, capital investment, debt, and market share, among others [5]–[7]. Regression models have been traditionally used to model changes in the stock markets. However, those models can predict linear patterns only [8]. The dominant data mining technique used in stock market prediction so far is neural network modeling, including back-propagation (BP) networks, probabilistic neural networks, and recurrent neural networks [9], [10]. The basic assumption is that similar input time series should produce similar output time series while ignoring intra-day fluctuations. Refenes *et al.* [11] compared regression models with a back-propagation network using the same data for stock prediction. Results showed that back-propagation network was a better predictor.

There are several important design issues involved in applying neural network approach to stock prediction: 1) determine the optimal length of time in the past from which to analyze data. Many studies take an aggregate of insider activities one month before the current date and then predict the future trend; 2) select time-sensitive indicators as network inputs; and 3) decide what to do with the lagged data. In general, the inputs to neural networks include daily transaction volume, interest rates, stock prices, moving average, and/or rate of change, etc. [5], [7], [12]. The above time-series data can be incorporated into neural networks as inputs in different ways.

- One can take the lagged data up to a number of weeks plus the current week, and use principle component analysis to form new inputs. As a result, an input pattern can consist of k past data points, $\mathrm{X(t)}, \mathrm{X(t-1)}, \ldots \mathrm{X(t-k)}$, and the output is for time $\mathrm{t}+1$. An alternative input is to use lagged differenced time-series data such as $\mathrm{X(t)-X(t-1)}, \mathrm{X(t-1)-X(t-2)}, \ldots \mathrm{X(t-k)-X(t-k-1)}$ [7].
- An alternative approach is to use copy-back/context units to integrate previous patterns into a later input pattern. For example, Wang and Leu [9] developed a recurrent BP neural network for forecasting mid-term price trend of Taiwan stock market. The network was trained using features extracted from AutoRegressive Integrated Moving Average (ARIMA) analyses. During training, the network fed back the difference of two previous successive predictions to the input layer in order to adjust connection weights.
- In view of recency and fluctuation of time-series data, weighted moving averages (the farther away the time is from the current date, the less weight values they carry) can provide a good prediction [10].

Typically, the performance of neural networks in classification problems such as stock prediction is measured by prediction accuracy: $\sum P_k/N$, where $p_k = 1$ (or 0) when predicted trend matches (or does not match) the actual trend. N is the total number of test patterns. Some researchers, however, point out that accuracy maximization may not be an appropriate goal for many real-world classification tasks from which natural data sets are taken. Classification accuracy assumes the known class distribution and equal misclassification costs for false positive and false negative errors [13]. In financial applications such as stock prediction and credit card fraud detection, however, one type of classification error is much more expensive than the other one. For example, the cost of incorrectly predicting a stock to be increasing or missing a case of fraud can be much more expensive than the cost of a false alarm. Therefore, some alternative metrics (e.g., using Receiver Operating Characteristic analysis) have been proposed [13].

Most neural network models that attempt to predict individual stocks only use information from the respective markets. Some studies attempt to include not only the current stock index value, but also trade volumes from all indexes in the neural network models to estimate the effect of more established markets whose values affect the performance of smaller emerging markets [7]. The assumption is that index forecasting performance will steadily increase with the inclusion of relevant external market indicators. In a study to predict a five-day future value for several market indexes [7], the input values in each BP neural network (each for a specific market) were the set of one-, two-, and five-day lags of the closing value, along with the corresponding one-, two-, and five-day normalized average trading volumes for the respective index markets. The only output value indicated a five-day future index value for the respective market. Results demonstrated that prediction performance was improved as additional external knowledge was added to the neural network.

Other methods such as rule induction, statistical analysis, genetic algorithm, and data visualization have also been used in stock prediction [14]–[16]. The Recon system, for example, induces classification rules to model the given data [14]. It analyzes a historical database and produces rules that will classify present stocks as either exceptional or unexceptional future performers. Each rule has its strength, and its prediction is weighted

by the amount of evidence supporting the rule. The rule-induction algorithm starts with distinguishing all numeric features. Then, it explores the space of all possible rules. The search through the rule space is essentially only a search through the space of the multivariate statistics of the data that is pertinent to the classification problem. It is reported that the Recon system has significantly outperformed the benchmark in terms of total return in a period of four years, indicating that rule induction is a valuable tool for stock selection. Some other approaches to predicting stock trends have combined ARIMA models and neural networks [9], [17]. An ARIMA model is a linear nonstationary model that uses difference operator to convert nonstationary series to stationary. It does not work well in modeling nonlinear series by itself.

Ankerst [18] used data visualization techniques to analyze the stock prices for Dow Jones, Gold, and IBM, etc. The individual vertical bars in the images corresponded to different years and the subdivision of the bars to the 12 months within each year. The coloring mapped high stock prices to light colors and low stock prices to dark colors, enabling the user to see the price change of a particular stock in a year easily.

A measure of the extent to which the growth rates are mixed provides an indication of uncertainty in the projected performance of the stock. Boston [6] proposed a measure based on a fuzzy logic model, which defined a measure of uncertainty by combining the growth rates of total revenues and earnings per share. An exploratory analysis of this uncertainty measure was conducted to predict returns on the stocks. The uncertainty measure of a stock was also anticipated to correlate with the variance in daily returns of that stock. The evaluation results revealed that stocks with low uncertainty were those with consistently high or consistently low rates. Stocks with high uncertainty were those with mixed growth rates, and their returns were expected to be moderate.

B. Portfolio Management

Portfolio management is a major issue in investment. It concerns how individuals decide which securities to hold in investment portfolios and how funds should be allocated among broader asset classes, such as stocks versus bonds, and domestic securities versus foreign securities. The primary goal is to choose a set of risk assets to create a portfolio in order to maximize the return under certain risk or to minimize the risk for obtaining a specific return. It is critical for investment models to have the best forecast of expected returns and risks in order to support financial decision-making. The definition of risk used in the classical Markowitz "Mean-Variance" model for effective portfolio is a measure of the variability of return called the standard deviation of return.

A number of adaptive supervised learning networks have been developed for traders and portfolio management [19]. LBS Capital Management Inc.[1], a fund-management firm, uses neural networks, genetic algorithms, and expert systems to manage portfolios worth US$600 million. The system has outperformed the overall stock market.

[1][Online] Available: http://www.lbs.com/.

The factor models such as the Capital Asset Pricing Model (CAPM) and Arbitrage Pricing Theory (APT) have been useful in explaining how assets are priced in the financial markets by relating their performance to their risk. Although these models do not provide any prediction of future asset returns, they offer important information for risk management. The most popular model in portfolio management in recent years is the APT [20]. The major assumption of the APT model is that the investment risks can be broken down into systematic and idiosyncratic risks. The APT approach suggests using quadratic programming to obtain the maximum of asset pricing. It helps identify a few risk factors from a set of candidates and select securities based on their relative risks and returns. APT has been integrated with neural networks in the portfolio management. In such a hybrid approach, an APT model can be used to determine prices, and then a neural network predicts the trend of each risk factor in the future. After investment alternatives are generated, the optimal portfolio is selected based on the investor's preference [21]. [22] designed a hybrid model for portfolio selection and management, which comprised three modules: a genetic algorithm for the selection of the assets that are going to form the investment portfolio, a neural net for the prediction of the returns on the assets in the portfolio, and a genetic algorithm for the determination of the optimal weights for each asset.

Sensitivity analysis of asset returns to various economic variables provides investors with a useful tool to build portfolios and manage their risk. Bentz *et al.* [23] compared two techniques for estimating asset future sensitivities to economic factors. The first technique was a Kalman filter to model the underlying dynamics of the sensitive coefficients. The second was a multilayer neural network trained with the standard BP algorithm. Assessment suggested that Kalman filter be only more effective than the neural network when the underlying sensitivities evolve gradually. Therefore, a method that combines both techniques in a single framework in order to model the time structure and conditional influences on the underlying sensitivities seems to be more effective.

Input-output hidden Markov models (IOHMMs) are conditional hidden Markov models in which the emission (and possibly the transition) probabilities can be conditioned on an input sequence. These conditional distributions can be linear, logistic, or nonlinear. Thus, IOHMMs have been used in financial time-series prediction tasks as well. Findings from some preliminary research have demonstrated that when predicting the conditional density of returns of market and sector indices, IOHMMs can yield significantly better performance, as estimated by the out-of-sample likelihood, than estimates with the historical average [24].

Data visualization techniques are primarily used to monitor the change and trend of individual investments in a portfolio [25]. Xiong *et al.* [26] applied ER modeling to visualize large mutual fund data and recommended a visualization approach for demonstrating the influence of a particular security on a fund's portfolio.

C. Bankruptcy Prediction

Predicting bankruptcy is of great benefit to those who have some relations to a firm concerned, for bankruptcy is a final state

of corporate failure. In the 21st Century, corporate bankruptcy in the world has reached an unprecedented level. It results in huge economic losses to companies, stockholders, employees, and customers, together with tremendous social and economical cost to the nation. Therefore, accurate prediction of bankruptcy has become an important issue in finance. Companies are strongly demanding explanations for the logic of prediction. They find it more acceptable to hear, for instance, that the prediction is produced based on computer-generated rules than to hear that the decision is made by an advanced technique that offers no explanation.

The breakthrough bankruptcy prediction model was the Z-score model developed by Altman [27]. The five-variable Z-score model using multiple discriminant analysis showed very strong predictive power. Since then, the discriminant analysis has been approved to be the most widely accepted and successful method in bankruptcy prediction literature [28]. In addition, numerous studies have tried to develop different bankruptcy prediction models by applying other data mining techniques including logistic regression analysis, genetic algorithms, decision trees, classification and regression trees (CART), and other statistical methods [29]–[31]. Those techniques can generally provide good interpretability of the prediction models. In the past two decades, a number of studies have also applied neural network approach to bankruptcy prediction, most centering on the comparison of predictive performance of neural networks and other methodologies such as discriminant analysis and logic analysis. Some have reported that the performance of neural networks is slightly better than that of other techniques, but results are contradictory or inconclusive [28].

Most bankruptcy models assume "normal" economic conditions, so they may be useless under "crisis" conditions. It has been found that data collected in a crisis are more influenced by noise than data collected under normal conditions. In Sung *et al.*'s study [28], the final input variables were 40 financial ratios categorized as growth, profitability, safety/leverage, activity/efficiency, and productivity. The multivariate discriminant analysis was used as the performance benchmark. After stepwise procedures under normal conditions, three variables were selected: cash flow to total assets, productivity of capital, and average turnover period for inventories. Under a crisis condition, three variables were also selected as the results of the same stepwise procedure: cash flow to liabilities, productivity of capital, and fixed assets to stockholders' equity and long-term liabilities. When the normal model was applied in crisis situations, accuracy of bankruptcy prediction dropped significantly, which justified the need for a different model in crisis economic conditions.

Although neural networks and statistical models have been used for bankruptcy prediction, they may encounter the problem of unequal frequencies of the two states of interest, which creates at least two major obstacles in evaluating the network predictive performance. The first issue involves the impact of unequal frequencies of the two states (e.g., bankruptcy versus not bankruptcy) on training a neural network or estimating the parameters of statistical models. Drawing random samples from unbalanced populations will likely yield samples that contain an overwhelming majority of one state of interest. Consequently, the decision performance of neural networks or statistical models may be poor while being tested in realistic situations. To overcome this problem, researchers have selected choice-based sampling technique in which the probability of an observation entering the sample depends on the value of the dependent variable. The second problem involves evaluating the accuracy of various decision models. The percentage of observations correctly classified can be very misleading with unbalanced samples [32]. In general, training a neural network with balanced samples in applications such as bankruptcy prediction can enable the network to familiarize itself with the infrequent state of interest. Neural networks trained on balanced samples provide the best results while being tested under realistic conditions. Jain and Nag [32] constructed several training samples with different composition. They compared the performance of a neural network that was trained on a balanced sample and the performance of another neural network trained on more representative samples. The weighted efficiency measure was the highest for the former network and decreased when the networks were trained using samples representative of the population.

D. Foreign Exchange Market

"Foreign Exchange" is the simultaneous buying of one currency and selling of another. The foreign exchange market is the largest financial market in the world, with a daily average turnover of over US$1 trillion. Currency traders make decisions using both technical factors and economic fundamentals. They have been finding that technical trading rules make excess returns in the foreign exchange market.

Data mining has been applied to identifying such technical trading rules. Some studies use genetic algorithms to find technical trading rules or simulate the volatility of exchange rate by evaluating and updating the mix of rules [33]–[35]. Walczak [36] used homogeneous neural network forecasting models for trading U.S. dollar against other foreign currencies. The neural networks were trained using more than 21 years of data to predict 1-day future spot rates for several nominal exchange rates, and achieved 58% accuracy for trading the British Pound and 57% accuracy for trading the German Mark. The input to each network was one or more spot rate lags. A k-day lag was calculated as $\mathrm{Lag_k} = \mathrm{X_t} - \mathrm{X_{t-k}}$, where $\mathrm{X_t}$ represented the currency exchange spot rate value at time t. The forecast value (output) for each neural network was the one-day future lag at time $\mathrm{t}+1$. Interestingly, the study also reported that once an initial minimal amount of training data produced the best forecasting performance, addition of more training data would not improve, and may even decrease, the performance of neural networks. Other researchers, however, have suggested that larger training sets produce better forecasting models [37].

It is well known that purchasing power parity and trade balance are two fundamental factors influencing the long-term movements of exchange rates. Traders and private individuals anticipate the direction of financial market movements before making an investment decision. In recent years, some researchers have attempted to investigate how money market

news headlines can be used to forecast intra-day currency exchange rate movements [38]. The innovation of this approach is that, unlike the analysis based on quantitative data, the forecasts are produced based on text describing the current status of financial markets, as well as political and general economic news (text mining). Such text contains not only the effect (e.g., rises), but also the possible causes. Peramunetilleke and Wong [38] proposed a text mining method to forecast short-term movements in the foreign exchange market from real-time news headlines and quoted exchange rates. In this approach, each news headline was associated with a time stamp showing the day, hour, and minute it was received through a news service, while the actual currency movements and time were extracted from time series of quoted exchange rates. The forecast was done as follows: 1) counting the number of occurrences of certain keywords in the news of each time period; 2) transforming the occurrences of keywords into weights (real values between 0 and 1). As a result, each keyword gets a weight value for each time period; 3) generating classification rules (expressing the correlation between the keywords and the outcome) based on weights and the closing values of the training data; and 4) applying generated rules to the news of the two most recent periods to yield the prediction.

E. Fraud Detection

Credit card transactions continue to grow, taking an ever-larger share of the U.S. payment system and leading to a higher rate of stolen account numbers and subsequent losses by banks. According to Meridien Research, financial institutions lost more than US$1 billion in credit and debit card fraud in 2001. Therefore, fraud detection is becoming a central application area of data mining, which aims at searching for patterns indicative of fraud. Improving fraud detection is essential to reducing the loss and maintaining the viability of the payment system.

Credit card fraud detection has highly peculiar characteristics. The first one is the very limited time span in which the acceptance or rejection decision has to be made. The second is that data are highly skewed: many more transactions are legitimate rather than fraudulent (otherwise, the entire industry would have soon ended up being out of businesses). The third characteristic is the huge amount of credit card operations that have to be processed at a given time. Although very few will be fraudulent, but this just means that the haystack where these needles are to be found is simply enormous.

A major task in fraud detection is the construction of algorithms or models that can learn how to recognize a variety of fraud patterns. These algorithms or models of fraudulent behavior can serve in decision support systems for preventing frauds or planning audit strategies [39]. HNC Software [40] uses a BP neural network with three layers in its system for detecting credit card fraud. The input layer receives information from an external source, the hidden layer uncovers common features of the information that the input layer provides, and the output layer learns combinations of the features that are associated with known behaviors or decisions.

Most data mining methods discard outliers as noise or exceptions. However, in fraud detection, the rare events can be more interesting than the more regularly occurring ones. Thus, outlier analysis has been used to detect fraudulent patterns that are substantially different from the main characteristics of regular credit card transactions [2]. Merix fraud detection system [41] incorporates expert systems, neural networks, and statistical methods, along with rough sets techniques, enhanced classifier systems, and advanced fuzzy logic rule induction systems, to provide customers with the ultimate fraud detection product. Stolfo *et al.* [42] combined local fraud detection agents with meta-learning system. Meta-learning is a general strategy that provides the means of learning how to combine and integrate a number of separately learned classifiers or models such as ID3, CART, and Bayes Classifier. The fraud-detection task exposes various technical problems including skewed distribution of training data, nonuniform cost per error, and scalability and efficiency issues (there are millions of transactions processed each day). One approach to addressing the efficiency and scalability issues is to divide a large data set of labeled transactions (either fraudulent or legitimate) into smaller subsets, then apply data mining techniques to generate multiple classifiers in parallel, and combine the resultant base models to generate a meta-classifier by learning from classifiers' behavior.

TABLE II
FINANCIAL APPLICATIONS AND RELEVANT DATA MINING TECHNIQUES

	Neural network	Genetic algorithm	Statistical inference	Rule induction	Data visualization
Prediction of stock market	[7, 9-12, 15, 43]	[16, 44, 45]	[11, 17, 43]	[14, 46, 47]	[48]
Portfolio management	[19, 21, 23]	[22, 49]	[24]	[50]	[25, 26]
Bankruptcy prediction	[28, 32, 34, 51, 52]	[30, 31]	[28, 30]	[29]	[53]
Foreign exchange market	[36, 37]	[33-35]	[54, 55]	[38]	[56]
Fraud detection	[40, 57-59]	[44]	[2]	[39, 41, 42]	[60]

Table II summarizes some of the past research on using data mining approaches in financial application. In general, among the five data mining techniques, neural network modeling is the most widely used approach to modeling the behavior and forecasting future values of financial time series [36]. In addition, statistics, genetic algorithms, and rule induction methods also play important roles in the financial services industry [61]–[63]. Data visualization techniques have yet been frequently used in financial applications due to the nature and complexity of data and problems [25]. In addition, visualization is incapable of automatically generating new knowledge.

F. Data Mining in Other Financial Applications

In addition to the above applications that we have discussed, data mining techniques have also been applied to other financial applications such as loan risk analysis and payment prediction [64], [65], mortgage scoring [66], [67], and real estate services.

Data mining systems can determine whether or not a customer will be able to pay off their loans based on his/her income, age, and historical credit information, etc. This is done by comparing the current customer information with historical data and categorizing the customer into one of the pre-defined

customer segments for loan payments, thus decreasing the risk for banks. Neural networks have been used for providing recommendations to grant or deny a loan based on financial ratios, past credit ratings, and loan records [65].

Currently, there are two prevalent technologies in mortgage scoring: logistic regression and neural networks. [67] used neural networks in the modeling of foreclosure of commercial mortgages. The study employed a large set of individual loan histories and showed that the neural network approach outperformed the logistic approach in terms of distinguishing between "good" and "bad" loans.

RealNegotiate[2] announces a data-mining software for real estate. It uses data mining techniques to analyze real market data (from any MLS) and helps real estate buyers, sellers, and professionals (realtors, investors, appraisers) answer questions such as "How soon will it be sold?", "How much faster will it be sold if we reduce the price by $5000?", "How much should I offer to get this property?", and "What are the odds of buying at this price?".

V. 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*— Knowledge discovery through data mining is an iterative process. The selection of appropriate variables and data mining algorithms, and model assessment and refinement are key components of this process. They should take features of financial data into account. Although neural network modeling is the most widely used method in data mining applications in finance, the optimal design of neural networks for various financial engineering problems remains open. A formal methodology for determining minimum/optimal training set size is required. Another important design issue involves finding the optimal length of time in the past for analysis and determining what to do with the lagged time-series data.
- *Scalability and performance*—Financial data are accumulated at an unprecedented pace. Data mining process must meet the challenges of scalability and computation efficiency. A large data set can be divided into multiple small ones, so that we can apply data mining techniques to those small data sets in parallel and integrate their results. In addition, real-life data are changing constantly, leading to the issue of model maintenance, which is concerned with the possible model change due to incremental updates of data. Incremental data mining has shown great potentials in addressing the above issue. It starts with a single model, and then updates the model with additional new data without creating a new model from scratch. The core of an incremental data mining algorithm is to determine whether or not a new data item falls into an existing category or fits into an existing model. Several algorithms have been explored, such as the grid-based incremental clustering algorithm and incremental-learning algorithms for neural networks [68].
- *Unbalanced frequencies of financial data*—Financial applications such as bankruptcy prediction and fraud detection are usually characterized by unequal frequencies of two states of interest. The presence of unequal frequencies can create at least two major obstacles in evaluating the performance of decision models, as discussed earlier in this paper. Therefore, how to handle unbalanced training data sets is critical to generating good financial models.
- *Text mining*—Most existing financial prediction systems are purely based on quantitative data such as stocks prices and market indices. In fact, textual articles appearing in leading financial newspapers such as the Wall Street Journal and Financial Times contain both effects (e.g., "stocks plummet", "rise") and experts' analyses, which make them an attractive resource for mining financial knowledge. A piece of news such as "the inflation rate is expected to increase next week" will probably cause an immediate movement in stock market shares.

  Natural language processing, information extraction, and text mining techniques can help make sense of textual information that is needed to support decision-making process. Research suggests that at least 80% of today's data are in an unstructured textual format. Text mining identifies patterns and predicts outcomes from large volumes of textual data [69]. A few financial text mining systems have attempted to use information contained in online articles to predict stock markets [38], [47]. Typically, those systems make use of one or more collections of keywords (e.g., "stock rally", "buy", "sell" and "bond strong") and weights or occurrence frequencies of each keyword in articles to generate probabilistic rules.
- *Mobile Finance*—With the latest advances in mobile computing and increasing bandwidth of wireless networks, more and more financial businesses are extended to the mobile environments. Several mobile data mining systems such as MobiMine have been developed to allow intelligent monitoring of time-critical financial data from a hand-held PDA. They facilitate stock monitoring and help detect the causal relationships between stock trends and different features characterizing the stocks. Basically, these systems make use of a collection of online mining techniques including statistical analysis, clustering, Bayesian networks, and decision trees.
- *Integration of multiple data mining techniques*—Any single data mining technique has its strengths and limitations. 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 [22], [70]. Wu *et al.* [70] have developed a hybrid system combining neural networks and fuzzy logic to forecast stock market performance. It is found 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.

[2][Online] Available: http://www.realnegotiate.com/.

- *Heterogeneous and distributed data sources*—As the globalization trend continues, financial data are widely distributed in various formats. The capability of dealing with disparate formats and distributed nature of data is indispensable to data mining applications. Techniques for data fusion and conflict resolution are helpful for bridging the gap among different data sources.

In this paper, we have discussed data mining in the context of financial application. Although data mining has been applied to 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] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI Mag.*, vol. 17, pp. 37–54, 1996.

[2] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*. San Francisco, CA: Morgan Kaufmann, 2001.

[3] W. Dwinnell, "Data visualization tips for data mining: pattern recognition provides data insight," *PC AI Mag.*, vol. 16.1, pp. 51–57, 2002.

[4] S. Taylor, *Modeling Financial Time Series*. New York: Wiley, 1986.

[5] A. N. Refenes, A. D. Zapranis, and Y. Bentz, "Modeling stock returns with neural networks," presented at the Workshop on Neural Network Applications and Tools, London, U.K., 1993.

[6] J. R. Boston, "A measure of uncertainty for stock performance," presented at the IEEE/IAFE/INFORMS 1998 Conf. Computational Intelligence for Financial Engineering (CIFEr), New York, 1998.

[7] S. Walczak, "Gaining competitive advantage for trading in emerging capital markets with neural networks," *J. Manag. Inform. Syst.*, vol. 16, pp. 177–192, 1999.

[8] J. Roman and A. Jameel, "Backpropagation and recurrent neural networks in financial analysis of multiple stock market returns," presented at the 29th HICSS, Wailea, HI, 1996.

[9] J.-H. Wang and J.-Y. Leu, "Stock market trend prediction using ARIMA-based neural networks," presented at the IEEE Int. Conf. Neural Networks, Washington, DC, 1996.

[10] A. M. Safer, "Predicting abnormal stock returns with a nonparametric nonlinear method," presented at the Int. Joint Conf. Neural Networks, Washington, DC, 2001.

[11] A. N. Refenes, A. D. Zapranis, and A. D. Francis, "Stock performance modeling using neural networks: a comparative study with regression models," *J. Neural Networks*, vol. 7, pp. 375–388, 1994.

[12] C. Ornes and J. Sklansky, "A neural network that explains as well as predicts financial market behavior," , Proc. IEEE IAFE, New York, Mar. 24–25, 1997, pp. 43–49.

[13] F. Provost, T. Fawcett, and R. Kohavi, "The case against accuracy estimation for comparing induction algorithms," presented at the 15th Int. Conf. Machine Learning, San Francisco, CA, 1998.

[14] G. H. John, P. Miller, and R. Kerber, "Stock selection using rule induction," *IEEE Expert*, vol. 11, pp. 52–58, 1996.

[15] E. W. Saad, D. V. Prokhorov, and D. C. Wunsch II, "Comparative study of stock trend prediction using time delay, recurrent and probabilistic neural networks," *IEEE Trans. Neural Networks*, vol. 9, pp. 1456–1470, Nov. 1998.

[16] M. A. Kaboudan, "Genetic programming prediction of stock prices," *Comput. Econ.*, vol. 16, pp. 207–236, 2000.

[17] M. Schumann and T. Lohrbach, "Comparing artificial neural networks with statistical methods within the field of stock market prediction," presented at the 26th HICSS, Wailea, HI, 1993.

[18] M. Ankerst, "Visual data mining with pixel-oriented visualization techniques," presented at the ACM SIGKDD Workshop on Visual Data Mining, San Francisco, CA, 2001.

[19] L. Xu and Y.-M. Cheung, "Adaptive supervised learning decision networks for traders and portfolios," presented at the IEEE/IAFE (Computational Intelligence for Financial Engineering), New York, 1997.

[20] S. Ross, "The arbitrage theory of capital asset pricing," *J. Econ. Theory*, vol. 13, pp. 341–360, 1976.

[21] S.-Y. Hung, T.-P. Liang, and V. W.-C. Liu, "Integrating arbitrage pricing theory and artificial neural networks to support portfolio management," *Decision Support Syst.*, vol. 18, pp. 301–316, 1996.

[22] J. G. Lazo, M. Maria, R. Vellasco, M. Aurélio, and C. Pacheco, "A hybrid genetic-neural system for portfolio selection and management," presented at the 7th Int. Conf. Engineering Applications of Neural Networks (EANN2000), Kingston Upon Thames, U.K., 2000.

[23] Y. Bentz, L. Boone, and J. Connor, "Modeling stock return sensitivities to economic factors with the Kalman filter and neural networks," presented at the IEEE/IAFE 1996 Conf. Computational Intelligence for Financial Engineering, Paris, France, 1996.

[24] Y. Bengio, V.-P. Lauzon, and R. Ducharme, "Experiments on the application of IOHMM's to model financial returns series," *IEEE Trans. Neural Networks*, vol. 12, pp. 113–123, Jan. 2001.

[25] LFE, "Financial Visualization," Lab. Financial Eng., MIT Sloan Rep., 2003.

[26] F. Xiong, E. C. Prakash, and K. W. Ho, "E-R modeling and visualization of large mutual fund data," *J. Visualiz.*, vol. 5, no. 2, pp. 197–204, 2002.

[27] E. Altman, "Financial ratios, discriminant analysis, and the prediction of corporate bankruptcy," *J. Finance*, vol. 23, pp. 589–609, 1968.

[28] T. K. Sung, N. Chang, and G. Lee, "Dynamics of modeling in data mining: interpretive approach to bankruptcy prediction," *J. Manag. Inform. Syst.*, vol. 16, pp. 63–85, 1999.

[29] C. Y. Shirata, "Peculiar behavior of Japanese bankrupt firms: discovered by AI-based data mining technique," presented at the 4th Int. Conf. Knowledge-Based Intelligent Engineering Systems and Allied Technologies, Tokyo, Japan, 2000.

[30] B. Back, T. Laitinen, K. Sere, and M. V. Wezel, "Choosing bankruptcy predictors using discriminant analysis, logit analysis, and genetic algorithms," Turku Center Comput. Sci., Finland, Tech. Rep. 40, Sept. 1996.

[31] K. S. Shin, "A genetic algorithm application in bankruptcy prediction modeling," *Int. J. Expert Syst. With Applic.*, vol. 23, pp. 321–328, 2002.

[32] B. A. Jain and B. N. Nag, "Performance evaluation of neural network decision models," *J. Manag. Inform. Syst.*, vol. 14, pp. 201–216, 1997.

[33] C. Neely, P. Weller, and R. Dittmar, "Is technical analysis in the foreign exchange market profitable? A genetic programming approach," *J. Financial Quant. Anal.*, vol. 32, pp. 405–426, 1997.

[34] F. Westerhoff and C. Lawrenz. (2000) Explaining exchange rate volatility with a genetic algorithm. Computing in Economics and Finance. [Online] Available: http://econpapers.hhs.se/paper/scescecf0/325.htm

[35] J. Arifovic, "The behavior of the exchange rate in the genetic algorithm and experimental economies," *J. Polit. Economy*, vol. 104, no. 3, pp. 510–541, 1996.

[36] S. Walczak, "An empirical analysis of data requirements for financial forecasting with neural networks," *J. Manag. Inform. Syst.*, vol. 17, pp. 203–222, 2001.

[37] G. Zhang and M. Y. Hu, "Neural network forecasting of the British pound/US dollar exchange rate," *Int. J. Manag. Sci.*, vol. 26, pp. 495–506, 1998.

[38] D. Peramunetilleke and R. K. Wong, "Currency exchange rate forecasting from news headlines," presented at the 13th Australasian Database Conf. (ADC 2002), Melbourne, Australia, 2002.

[39] F. Bonchi, F. Giannotti, G. Mainetto, and D. Pedreschi, "A classification-based methodology for planning audit strategies in fraud detection," presented at the 5th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, San Diego, CA, 1999.

[40] M. Weatherford, "Mining for fraud," *IEEE Intell. Syst.*, vol. 17, pp. 4–6, Jan./Feb. 2002.

[41] Merix, Fraud Detection, Nutech Solution, Inc., Charlotte, NC.

[42] S. J. Stolfo, W. D. Fan, W. Lee, A. Prodromidis, and P. Chan, "Credit card fraud detection using meta-learning issues and initial results," in *Proc. AAAI-97 Workshop on Al methods in fraud and risk management*, Menlo Park, CA, July 1997, pp. 83–90.

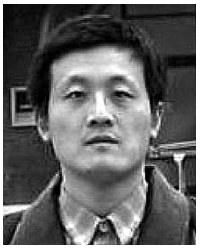
[43] Y. Yoon, G. Swales, and T. M. Margavio, "A comparison of discriminant-analysis versus artificial neural networks," *J. Oper. Res. Soc.*, vol. 44, pp. 51–60, 1993.

[44] H. Iba and T. Sasaki, "Using genetic programming to predict financial data," presented at the 1999 Congr. Evolutionary Computation, Washington, DC, 1999.

[45] ——, "Using genetic programming to predict financial data," in *IEEE Congr. Evolutionary Computation*, Washington, DC, 1999.

[46] T. Chenoweth and Z. Obradovic, "A multicomponent nonlinear prediction system for the S&P 500 index," *Neurocomputing*, vol. 10, pp. 275–290, 1996.

[47] B. Wuthrich, V. Cho, S. Leung, D. Permunetilleke, K. Sankaran, J. Zhang, and W. Lam, "Daily stock market forecast from textual web data," presented at the 1998 IEEE Int. Conf. Systems, Man, and Cybernetics, San Diego, CA, 1998.
[48] U. Fayyad, G. Grinstein, and A. Wierse, *Information Visualization in Data Mining and Knowledge Discovery*. San Francisco, CA: Morgan Kaufmann, 2001.
[49] R. M. Stein and R. N. Bernard, "Data mining the future: genetic discovery of good trading rules in agent-based financial market simulations," presented at the IEEE/IAFE/INFORMS 1998 Conf. Computational Intelligence for Financial Engineering, New York, 1998.
[50] G. H. John and P. Miller, "Building long/short portfolios using rule induction," presented at the IEEE Conf. Computational Intelligence in Financial Engineering, New York, 1996.
[51] R. L. Wilson and R. Sharda, "Bankruptcy prediction using neural networks," *Decision Support Syst.*, vol. 11, pp. 545–557, 1994.
[52] E. Rahimian, S. Singh, T. Thammachote, and R. Virmani, "Bankruptcy prediction by neural network," in *Neural Networks in Finance and Investing*, R. R. Trippi and E. Turban, Eds. New York: McGraw-Hill, 1993, pp. 159–176.
[53] K. Kiviluoto and P. Bergius, "Exploring corporate bankruptcy with two-level self-organizing maps. Decision technologies for computational management science," presented at the 5th Int. Conf. Computational Finance, Boston, MA, 1998.
[54] H. Tsurumi and T. Nakatsuma, "Bayesian analysis of doubly truncated regression models with ARMA-GARCH errors with an application to stock returns and foreign exchange rates," presented at the 4th World Meeting Int. Soc. Bayesian Analysis, Cape Town, South Africa, 1996.
[55] F.-M. Tseng, G.-H. Tzeng, H.-C. Yu, and B. J. C. Yuan, "Fuzzy ARIMA model for forecasting the foreign exchange market," *Fuzzy Sets Syst.*, vol. 118, pp. 9–19, 2001.
[56] D. L. Gresh, B. E. Rogowitz, M. S. Tignor, and E. J. Mayland, "An interactive framework for visualizing foreign currency exchange," presented at the IEEE Visualization'99, San Francisco, CA, 1999.
[57] R. Brause, T. Langsdorf, and M. Hepp, "Neural data mining for credit card fraud detection," in *Proc. 11th IEEE Int. Conf. Tools with Artif. Intell.*, Chicago, IL, Nov. 8–10, 1999, pp. 103–106.
[58] S. Ghosh and D. L. Reilly, "Credit card fraud detection with a neural network," in *Proc. 27th Hawaii Int. Conf. Syst. Sci.*, vol. 3, Maui, HI, Jan. 4–7, 1994, pp. 621–630.
[59] J. R. Dorronsoro, F. Ginel, C. Sanchez, and C. S. Cruz, "Neural fraud detection in credit card operations," *IEEE Trans. Neural Networks*, vol. 8, pp. 827–834, July 1997.
[60] Viscovery. Viscovery for CRM-Applications. Eudaptics Software. [Online] Available: http://www.eudaptics.com
[61] W. B. Fairley, "Investment income and profit margins in property-liability insurance: theory and empirical results," *Bell J. Econ.*, vol. 10, pp. 192–210, 1979.
[62] L. F. Rozanova, R. D. Shagaliev, and Z. V. Maximenko, "Methods of quantitative analysis of real investment projects risks," presented at the 2nd Int. Workshop on Computer Science and Information Technologies (CSIT'2000), Ufa, Russia, 2000.
[63] A. S. Weigend and S. Shi, "Predicting daily probability distributions of S&P500 returns," *J. Forecasting*, vol. 19, pp. 375–392, 2000.
[64] R. Gerritsen, "Assessing loan risks: a data mining case study," *IT Professional*, vol. 1, pp. 16–21, 1999.
[65] R. P. Srivastava, "Automating judgmental decisions using neural networks: a model for processing business loan applications," presented at the ACM Annu. Conf. Communications, Kansas City, MO, 1992.
[66] G. H. John and Y. Zhao, "Mortgage data mining," in *Proc. IEEE/IAFE*, New York, Mar. 23–25, 1997, pp. 232–236.
[67] A. Episcopos, A. Pericli, and J. Hu, "Commercial mortgage default: a comparison of logit with radial basis function networks," *J. Real Estate Finance and Econ.*, vol. 17, pp. 163–178, 1998.
[68] R. Polikar, L. Udpa, S. S. Udpa, and V. Honavar, "Learn++: an incremental learning algorithm for supervised neural networks," *IEEE Trans. Syst., Man, Cybern. C*, vol. 31, pp. 497–508, 2001.
[69] N. Leavitt, "Data mining for the corporate masses?," *IEEE Comput.*, vol. 35, pp. 22–24, 2002.
[70] X. Wu, M. Fung, and A. Flitman, "Forecasting stock market performance using hybrid intelligent system," presented at the Int. Conf. Computational Science (ICCS), San Francisco, CA, 2001.



Dongsong Zhang received the Ph.D. degree in management from the University of Arizona, Tucson, in 2002.

Currently, he is an Assistant Professor in the Department of Information Systems, University of Maryland, Baltimore County. His research interest includes Web-based learning, data mining, computer-mediated communication, mobile computing, and Web service applications.

Lina Zhou received the Ph.D. degree in computer science from Beijing University, Beijing, China, in 1998.

She is an Assistant Professor in the Department of Information Systems, University of Maryland, Baltimore County. Her research interests center around text mining, deception detection, ontology learning, and machine translation.