SYNOPSIS ON
MINING FINANCIAL TIME SERIES DATABASES USING
MACHINE INTELLIGENCE AND EVOLUTIONARY
COMPUTING TECHNIQUES

(A Ph. D. Thesis Topic in Computer Science and Engineering)

Candidates Name: Sujata Chakravarty

1. INTRODUCTION AND LITERATURE SURVEY:

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 (Berka, 2002) 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 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. Almost every computational method has been explored and used for financial modeling. We will name just a few recent studies: Monte-Carlo simulation of option pricing, finite-difference approach to interest rate derivatives, and fast Fourier Transform for derivative pricing (Huang et. al., 2004; Zenios, 1999; Thulasiram and Thulasiraman, 2003). New developments augment traditional technical analysis of stock market curves (Murphy, 1999) that has been used extensively by financial institutions. Such stock charting helps to identify buy/sell signals (timing “flags”) using graphical patterns. Data mining as a process of discovering useful patterns, correlations has it own niche in financial modeling. Similarly to other computational methods almost every data mining method and technique has been used in financial modeling. An incomplete list includes a variety of linear and non-linear models, multi-layer neural networks (Kingdon, 1997; Walczak, 2001; Thulasiram et. al., 2002; Huang et. al., 2004) k-means and hierarchical clustering; k-nearest neighbors, decision tree analysis, regression, logistic regression, general multiple regression), ARIMA, principal component analysis, and Bayesian
learning. Less traditional methods used include rough sets (Shen and Loh, 2004), relational data mining methods (deterministic inductive logic programming and newer probabilistic methods (Muggleton, 2002; Lachiche and Flach, 2002; Kovalerchuk and Vityaev, 2000), support vector machine, independent component analysis, Markov models and hidden Markov models. Bootstrapping and other evaluation techniques have been extensively used for improving data mining results. Specifics of financial time series analyses with ARIMA, neural networks, relational methods, support vector machine and traditional technical analysis is discussed in (Back and Wigend, 1998; Kovalerchuk and Vitaasv, 2000; Muller et.al., 1997; Murphy, 1999; Tsay, 2002).

The naïve approach to data mining in finance assumes that somebody can provide a cookbook instruction on “how to achieve the best result”. Some publications continue to foster this unjustified belief. In fact, the realistic approach proven to be successful is providing comparisons between different methods showing their strengths and weaknesses relative to problem characteristics (problem ID) conceptually and leaving for user the selection of the method that likely fits the specific user problem circumstances. In essence this means clear understanding that data mining in general, and in finance specifically, is still more art than hard science. Fortunately, now there is growing number of books that discuss issues of matching tasks and methods in a regular way (Dhar and Stein, 1997; Kovaler, Chuck and Vitaev, 2002; Wang, 2003).

2. DATA MINING FOR FINANCIAL APPLICATIONS

For instance, understanding the power of first-order If-Then rules over the decision trees can significantly change and improve data mining design. User’s actual experiments with data provide a real judgment of data mining success in finance. In comparison with other fields such as geology or medicine, where test of the forecast is expensive, difficult, and even dangerous, a trading forecast can be tested next day in essence without cost and capital risk involved in real trading. Attribute-based learning methods such as neural networks, the nearest neighbors’ methods, and decision trees dominate in financial applications of data mining. These methods are relatively simple, efficient, and can handle noisy data. However, these methods have two serious drawbacks: a limited ability to represent background knowledge and the lack of complex relations. Relational data mining techniques that include Inductive Logic Programming (ILP) (Muggleton, 1999;
Dseroski, 2002) intend to overcome these limitations. Previously these methods have been relatively computationally inefficient (Thulasiram, 1999) and had rather limited facilities for handling numerical data (Bratko and Muggelton, 1995).

Various publications have estimated the use of data mining methods like hybrid architectures of neural networks with genetic algorithms, chaos theory, and fuzzy logic in finance. “Conservative estimates place about 5 billion to 10 billion under the direct management of neural network trading models. This amount is growing steadily as more firms experiment with and gain confidence in neural networks techniques and methods” (Loofbourrowant Loofbourrow, 1995). Many other proprietary financial applications of data mining exist, but are not reported publicly as was stated in (Von Atrock, 1997; Groth, 1998).

3. SPECIFICS OF DATA MINING IN FINANCE

Specifics of data mining in finance are coming from the need to: forecast multidimensional time series with high level of noise; accommodate specific efficiency criteria (e.g., the maximum of trading profit) in addition to prediction accuracy such as $R^2$; make coordinated multiresolution forecast (minutes, days weeks, months and years); incorporate a stream of text signals as input data for forecasting models (e.g. Enron case, September 11 and others). It has been shown that the financial data are not random and that the efficient market hypothesis is merely a subset of a larger chaotic market hypothesis (Drake and Kim, 1997). This hypothesis does not exclude successful short term forecasting models for prediction of chaotic time series (Casdagil and Eubank, 1992). Data mining does not try to accept or reject the efficient market theory. Data mining creates tools which can be useful for discovering subtle short-term conditional patterns and trends in wide range of financial data.

The impact of market players on market regularities stimulated a surge of attempts to use ideas of statistical physics in finance (Bouchaud and Potters, 2000). If an observer is a large marketplace player then such observer can potentially change regularities of the marketplace dynamically. Attempts to forecast in such dynamic environment with thousands active agents leads to much more complex models than traditional data mining models designed for. This is one of the major reasons that such interactions are modeled using ideas from statistical physics rather than from statistical data mining.
Data mining approach covers empirical models and regularities derived directly from data and almost only from data with little domain knowledge explicitly involved. Historically, in many domains, deep field-specific theories emerge after the field accumulates enough empirical regularity. We see that the future of data mining in finance would be to generate more empirical regularities and combine them with domain knowledge via generic analytical data mining approach (Mitchell, 1997). First attempts in this direction are presented in (Kovalerchuck and Vityaev, 2000) that exploit power of relational data mining as a mechanism that permits to encode domain knowledge in the first order.

4. TIME SERIES ANALYSIS

A temporal dataset $T$ called a time series is modeled in attempt to discover its main components such as Long term trend, $L(T)$, Cyclic variation, $C(T)$, seasonal variation $S(T)$ and irregular movements, $I(T)$: Assume that $T$ is a time series such as daily closing price of a share, or S&P 500 index from moment 0 to current moment $k$, then the next value of the time series $T(k + n)$ is modeled by formula:

$$T(k + n) = L(T) + C(T) + S(T) + I(T)$$  \hspace{1cm} (1.1)

Traditionally, classical ARIMA models occupy this area for finding parameters of functions used in formula 1.1. ARIMA models are well developed but are difficult to use for highly non-stationary stochastic processes. Potentially data mining methods can be used to build such models to overcome ARIMA limitations. The advantage of this four-component model in comparison with “black box” models such as neural networks is that components in formula 1.1 have an interpretation.

4.1 DATA SELECTION AND FORECAST HORIZON

Data mining in finance has the same challenge as general data mining in data selection for building models. In finance, this question is tightly connected to the selection of the target variable. There are several options for target variable $y$: $y = T(K + 1), T(K + 2), \ldots : y = T(K + n)$, where $T(K + 1)$ represents forecast for the next time moment, and $y = T(k + n)$ represents forecast for $n$ moments ahead. Selection of dataset $T$ and its size for a specific desired forecast horizon $n$ is a significant challenge. For stationary stochastic processes the answer is well-known a better model can be built for longer training duration. For financial time series such as S&P 500 index this is not the
case (Mehta and Bhattacharya, 2004). Longer training duration may produce many and contradictory profit patterns that reflect bear and bull market periods. Models built using too short durations may suffer from overfitting and hardly applicable to the situations where market is moving from the bull period to the bear period. Also in finance the long-horizon returns could be forecast better than short-horizon returns depending on the training data used and model parameters (krolzig et.al., 2004). In standard data mining it is typically assumed that the quality of the model does not depend of frequency of its use. In financial application the frequency of trading is one of parameters that impact a quality of the model. This happens because in finance the criterion of the model quality is not limited by the accuracy of prediction, but it driven by profitability of the model. It is obvious that frequency of trading impacts the profit as well as the trading rules and strategy.

5. ASPECTS OF DATA MINING METHODOLOGY IN FINANCE

Data mining in finance typically follows a set of general steps for any data mining task such as problem understanding, data collection and refining, building a model, model evaluation and deployment (Klosgen and Zytkow, 2002). Some specifics of these steps for trading tasks are presented in (Zemke, 2002; Zemke, 2002) such as data enhancing techniques, predictability tests, performance improvements, and pitfalls to avoid. Another important step in this process is adding expert-based rules in data mining loop when dealing with absent or sufficient data. “Expert mining” is a valuable additional source of regularities. However, in finance, expert-based learning systems respond slowly to the market changes (Cwan, 2002). A technique for efficiently mining regularities from an expert’s perspective has been offered (Kovalerchuk and Vityaev, 2000). Such techniques need to be integrated into financial data mining loop similar to what was done for medical data mining application (Kovalerchuck et. al., 2001).

5.1. ATTRIBUTE-BASED AND RELATIONAL METHODOLOGIES

Several parameters characterize data mining methodologies for financial forecasting. Data categories and mathematical algorithms are most important among them. The first data type is represented by attributes of objects, that is each object x is given by a set of values \( A_1(x), A_2(x), \ldots, A_n(x) \). The common data mining methodology assume this type
of data and it is known as *attribute-based or attribute-value* methodology. It covers a wide range of statistical and connectionist (neural network) methods.

The *relational data* type is second type, where objects are represented by their relations with other objects, for instance, \(x>y\), \(y<z\) and \(x>z\). In this example, we may not know that \(x=3\), \(y=1\) and \(z=2\). Thus attributes of objects are not known, but their relations are known. Objects may have different attributes (e.g. \(x=5\), \(y=2\) and \(z=4\)), but still have the same relations. Less traditional relational methodology is based on the relational data type. Another data characteristic important for financial modeling methodology is an actual *set of attributes* involved. A fundamental analysis approach incorporates all available attributes, but technical analysis approach is based only on a time series such as stock price and parameters derived from it. Most popular time series are index value at open, index value at close, highest index value, lowest index value and trading volume and lagged returns from the time series of interest. Fundamental factors include the price of gold, retail sales index, industrial production indices, and foreign currency exchange rates. Technical factors include variables that are derived from time series such as moving averages. The next characteristic of a specific data mining methodology is a form of the relationship between objects. Many data mining methods assume a *functional form* of the relationship. For instance, the linear discriminant analysis assumes linearity of the border that discriminates between two classes in the space of attributes. Often it is hard to justify such functional form in advance. Relational data mining methodology in finance does not assume a functional form for the relationship. Its intention is learning symbolic relations on numerical data of financial time series.

6. DATA MINING MODELS AND PRACTICE IN FINANCE

Prediction tasks in finance typically are posed in one of two forms: (1) straight prediction of the market numeric characteristic, e.g. stock return or exchange rate, and (2) the prediction whether the market characteristic will increase or decrease. Having in mind that we need to take into account the trading cost and significance of the trading return in the second case we need to forecast whether the market characteristic will increase or decrease no less than some threshold. Thus, the difference between data mining methods for (1) or (2) can be less obvious, because (2) may require some kind of numeric forecast.
6.1 Portfolio management and neural networks

The neural network most commonly used by financial institutions is a multilayer perceptron (MLP) with a single hidden layer of nodes for time series prediction. The peak of research activities in finance based on neural networks was in mid 1990s (Trippi and turban, 19996, Freedman et. al, 1995; Azoff, 1994) that covered MLP and recurrent NN (Rfenes, 1995). Other neural networks used in prediction are time delay networks, Elman networks, Jordan networks, GMDH, multi-recurrent networks (Giles et. al., 1997). This consideration also shows current challenges of data mining in finance- the need to build models that can be very quickly evaluated in both accuracy and interpretability. It is likely that extracting rules from the neural network is a temporary solution. It would be better to extract rules directly from data without introducing neural network artifacts to rules and potentially overlooking some better rules because of this. It is clear that it can happen from mathematical considerations.

There are also a growing number of computational experiments that support this claim, e.g., see (Kovalerchuk and Vityaev, 2000) on experiments with S&P 500, where first order rules built directly from data outperformed back propagation neural networks that are most common in financial applications. (Moody and Saffell, 2001) discuss advantages of incremental portfolio optimization and building trading models. The logic of using data mining in trading futures is similar to portfolio management.

The logic of portfolio management based on discovering interpretable trading rules is the same as for neural networks with the substitution of NN for rule discovering techniques.

7. OBJECTIVE OF THE PROPOSED RESEARCH

Financial prediction like bank prime loan, discount rate, federal funds, portfolios, prices of commodity markets like energy, etc. involve uncertainties in the nature of the data. Therefore interval type fuzzy sets will be used along with linear wavelet networks for financial prediction and knowledge discovery in data solely for the purpose of data mining. The interval type fuzzy sets have membership values in an interval over which the secondary membership degree is always 1. There are fast algorithms to compute the output of the interval type fuzzy sets for forecasting problems in the financial domain. A possible hybrid combination of interval type fuzzy sets and wavelet network will be explored for prediction of bank rates, stock market and currency fluctuations. Swarm
intelligence and differential evolution strategy will be adopted for optimizing the membership functions and the weights of the wavelet network. The interactive evolutionary strategy is expected to provide robust prediction in the face of uncertain fluctuations in the financial time series.

Electricity markets are playing major roles in buying and selling electricity either through interconnected markets or bilateral contracts. This is an important issue in the electricity market deregulation in many countries of the west. Given a wide variety of options in trading of electricity, it is essential to forecast the electricity market prices in an optimal way. It is envisaged to forecast one to 24 hours a head electricity prices taking into account the total electricity demand, outliers in the input dataset. Other types of time series databases in financial applications will be tried by variety of machine intelligence techniques including Support Vector Machines and Statistical Learning Theory.

8. References


