A Discrete Stock Price Prediction Engine Based on Financial News

Robert P. Schumaker[1] and Hsinchun Chen[2]

[1]Information Systems Dept.
Iona College, New Rochelle, New York 10801, USA
rschumaker@iona.edu

[2]Artificial Intelligence Lab, Department of Management Information Systems
The University of Arizona, Tucson, Arizona 85721, USA
hchen@eller.arizona.edu

Word Count: 3571

Abstract

This work investigates the possibility of discrete stock price prediction using a synthesis of linguistic, financial and statistical techniques to create the Arizona Financial Text System (AZFinText). In particular we compare AZFinText's predictions against existing quantitative funds using textual representation and statistical machine learning methods on financial news articles. Through our research, the AZFinText system managed a Simulated Trading return of 8.50% (compared to 5.62% for the S&P 500 index). In direct comparisons to existing quantitative mutual funds, our system's trading return performed well against the top 10 quantitative mutual funds of 2005, where our system would have placed fifth. When comparing AZFinText against only those quantitative funds that monitor the same securities, AZFinText managed a 2% higher return than the best performing quant fund.

Keywords

Decision Support, Text Analysis, SVM, Prediction, Quantitative Analysis

Introduction

The ability to predict stock market movement has been a source of interest for many researchers. While numerous scientific attempts have been made, no single method has yet been discovered to accurately predict stock price movement. Difficulty in prediction comes from the complexities associated with market dynamics where parameters are constantly shifting and are not fully defined.

One area of limited success in stock market prediction comes from textual data. Information from quarterly reports or breaking news stories can dramatically affect the share price of a security. Applying computational methods to this textual data forms the basis of financial text mining. Most existing literature on financial text mining applies a representational technique to news articles where only certain terms are used and weights are assigned to the terms based on the direction the stock price moves. Prediction then applies these weighted terms to a new article to determine a likely direction of movement. To their credit, these simpler forms of analyses have shown a weak but definite ability to predict price direction but not the price itself.

However, using computational approaches to predict stock prices using financial data is not unique. In recent years, interest has increased in Quantitative funds, or Quants, that automatically sift through numeric financial data and issue stock recommendations [5]. While these systems are based on proprietary technology, they do differ in the amount of trading control they have, ranging from simple stock recommenders to trade executors. Using historical market data and complex mathematical models, these methods are constrained to make assessments within the scope of existing information. This weakness means that they are unable to react to unexpected events falling outside of historical norms. However, this disadvantage has not stopped fund managers at Federated, Janus, Schwab and Vanguard from trusting billions of dollars of assets to the decisions of these computational systems.

In this paper, we introduce a different type of quant trader called AZFinText (Arizona Financial Text system), which focuses on making discrete numeric predictions based on the combination of financial news articles and stock price quotes. Our contribution rests on system building of the AZFinText system where trends and patterns are machine learned from stock quotes and textual financial news. While prior textual financial research has relied on tracking price direction alone, AZFinText leverages statistical learning to generate numeric price predictions and then make trading decisions from them. We further demonstrate that AZFinText outperforms the market average and performs well against existing quant funds.

Prediction of Securities

There are several theories concerning security forecasting, the first of which is the Efficient Market Hypothesis (EMH). Within the confines of EMH it is assumed that the price of a security is a direct reflection of all information available and that everyone has some degree of access to this information. From the principles of EMH, it is believed the markets are efficient and react instantaneously to new information by immediately incorporating it into the share price. A different perspective on prediction comes from Random Walk Theory where prices fluctuate randomly in the short-term. This theory has similar theoretical underpinnings to EMH where both believe that all public information is available to everyone and that consistently outperforming the market is an impossibility.

From these theories, two distinct trading philosophies emerged; the fundamentalists and the technicians. In a fundamentalist trading philosophy, the price of a security is determined through a myriad of financial numbers and ratios. Numbers such as inflation, return on equity and price to earnings ratios can all play a part in determining the price of a stock. Time-series data is not considered in a fundamental strategy but is a critical part of technical analysis. Technicians reason that price movements are not random and that patterns can be identified. However, technical analysis is considered to be an art form and as such is subject to interpretation. Researchers also believe that there is a window of opportunity where weak prediction exists before the market corrects itself to equilibrium. Using this small window of opportunity (in hours or minutes) and an automated textual news parsing system, the possibility exists to capitalize on stock price movements before human traders can act.

A. Algorithmic Quants

Among trading professionals there has been significant interest in the computational analysis of financial data. Their systems follow various stock parameters and are essentially automated versions of existing market strategies (e.g., look for high growth, unvalued securities, etc.) except with the ability to follow all stocks in real-time. This advantage has led quants to steadily outperform market averages by 2-3% for the past several years [5].

While the exact strategies used are a closely guarded secret, some quantitative funds do disclose the parameters they track. The exact number and weights assigned to these parameters fluctuate frequently to keep pace with market conditions and to tweak model performance. Quant programs are also becoming a part of the individual investor's toolbox as well. Wealth Lab Pro software allows individual investors to track upwards of 600 parameters through 1,000 pre-set investment strategies [7].

The number of quant funds has increased from just a few in 2001 to over 150 by the beginning of 2006 [2]. These funds have also branched themselves out, able to cover worldwide financial markets or focus exclusively on a select boutique of securities.

Quants generally operate in the following two-stage manner. First, securities are analyzed using a technical analysis strategy and securities not meeting basic criteria are removed from further analysis. Second, the quantitative algorithm rank orders the remaining stocks. Figure 1 illustrates a brief taxonomy of technical stock prediction algorithms used by quants as well as their discipline of origin. While this figure is no way an exhaustive list, we highlight some of the more important algorithms.

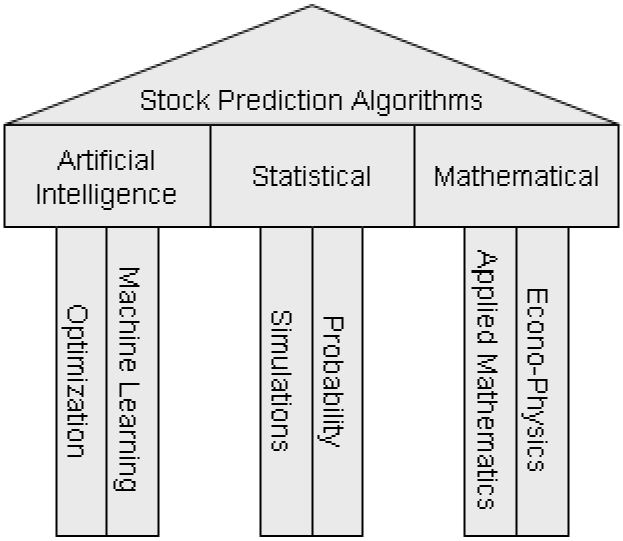


Figure 1. Technical Stock Prediction Algorithms

A-1. Artificial Intelligence

Artificial Intelligence has mostly contributed algorithms that deal with optimization and machine learning. Examples such as Genetic Algorithms, Support Vector Machines (SVM) and Neural Networks all take input parameters from financial securities and return predictions based on the hidden patterns within the historical data. However, most of these techniques have been constrained to either identifying the most relevant parameters or evaluating stock data in terms of a direction of movement. Examples include Genetic Algorithms which utilize a global search and optimization approach to identify the parameters that have the greatest impact on stock price performance [6]; SVM which is a machine learning algorithm that can classify the potential stock price into likely movement directions such as rise, drop or no recommendation; and Neural Networks which function by weighting various stock parameters. All of these methods performed marginally better than chance in prior research [4, 6, 14].

A-2. Statistical

Statistical approaches use simulations and probability methods such as Monte Carlo and Game Theory [3]. In Monte Carlo simulations, the problem of price prediction is too difficult to approach directly, so input parameters are given a series of suitable random numbers and are observed for how close they arrive at the predicted value [11].

In Game Theory, price prediction is modeled in terms of strategies and potential payoff. It is theorized that the players in the game will evaluate other player's strategies and adopt a stance which will earn them the best payoff. However, these types of systems do not function well in stock market prediction because of new entrants, constantly changing strategies from other players and the inherent difficulty in predicting price changes.

A-3. Mathematical

Mathematical approaches borrow heavily from the areas of applied mathematics and econo-physics. This branch of predictive algorithms uses more complex mathematical formalisms, such as Percolation Methods, Log-Periodic Oscillations and Wavelet Transforms to model future prices [11].

Percolation Methods use dimensional membranes to constrict trading actions and price movements. In one such example a lattice of traders is modeled, where a cluster of traders indicate a single company and at each time interval traders are given the choice to buy, sell, or sleep. This method is then used to model the supply and demand of securities and the potential impact on security prices.

Log-Periodic Oscillations use long-term historical data to describe macro movements in the market, such as impending crashes and market bubbles. While it has been suggested that previous ‘crash’ predictions from this model had more to do with luck, market psychology would make these oscillations more pronounced through rapid sell-offs in the face of an anticipated crash [11].

In Wavelet Transforms, input parameters are consecutively sampled to provide a finer-grained resolution into the microscopic movements that comprise the input signal. These successive filters can then be analyzed to provide parameter relation insights.

B. Financial News Representation

In order to address the weaknesses of current quant systems and obviate some of the risks associated with unexpected news, researchers have focused on learning patterns from financial news articles and making predictions from them [1]. The “Bag of Words” approach has emerged as a standard representation in textual financial research because of its ease of use. This process involves removing stopwords such as conjunctions and declaratives from the text and using what remains as the textual representation. While this method has been popular, it undergos noise-related issues associated with seldom-used terms and problems of scalability where immense computational power is required for large datasets. To improve on many of the representational and scalability problems, Noun Phrases retains only the nouns and noun phrases within a document and has been found to adequately represent the important article concepts. A third representational technique is Named Entities, which extends Noun Phrases by selecting the proper nouns of an article that fall within well-defined categories. This process uses a semantic lexical hierarchy as well as a syntactic/semantic tagging process to assign terms to categories. Selected categorical definitions are described by the Message Understanding Conference (MUC-7) Information Retrieval task and encompass the entities of date, location, money, organization, percentage, person and time. This method allows for better generalization of previously unseen terms and removes many of the scalability problems associated with a semantics-only approach. A fourth representational technique is Proper Nouns which functions as an intermediary between Noun Phrases and Named Entities. This representation is a subset of Noun Phrases and a superset of Named Entities, without the constraint of pre-defined categories. This representation removes the ambiguity associated with those particular proper nouns that could be represented by more than one named entity category or fall outside one of the seven defined Named Entity categories. In a comparison of different textual representation schemes, Bag of Words was found to be least effective in predicting future prices where the Proper Nouns of an article was most effective because of its concise nature to represent an article [9]. Another problem that arises in textual representation is infrequent term usage where a term may appear only one or two times in an entire corpus. Basing predictions off infrequent term appearances can lead to unpredictable results. To reduce the impact of infrequent term use, a cutoff is often introduced where only those terms that appear multiple times in any one article are used. This strategy effectively limits the number of text features in support of scalability.

Once financial news articles are represented, computers can then begin the task of identifying the patterns responsible for predictable behavior. One accepted method, Support Vector Regression (SVR), is a regression equivalent of Support Vector Machines (SVM) without the aspect of classification [13]. Like SVM, SVR attempts to minimize its fitting error while maximizing its goal function by fitting a regression estimate through a multi-dimensional hyperplane. This method is also well-suited to handling textual input as binary representations and has been used in similar financial news studies [10, 12].

Arizona Financial Text (AZFinText) System

In order to address the weaknesses of quants to unexpected information, we developed the AZFinText system. AZFinText is a machine learning system that uses financial news articles and stock quotes as its input parameters. Figure 2 illustrates the AZFinText system design.

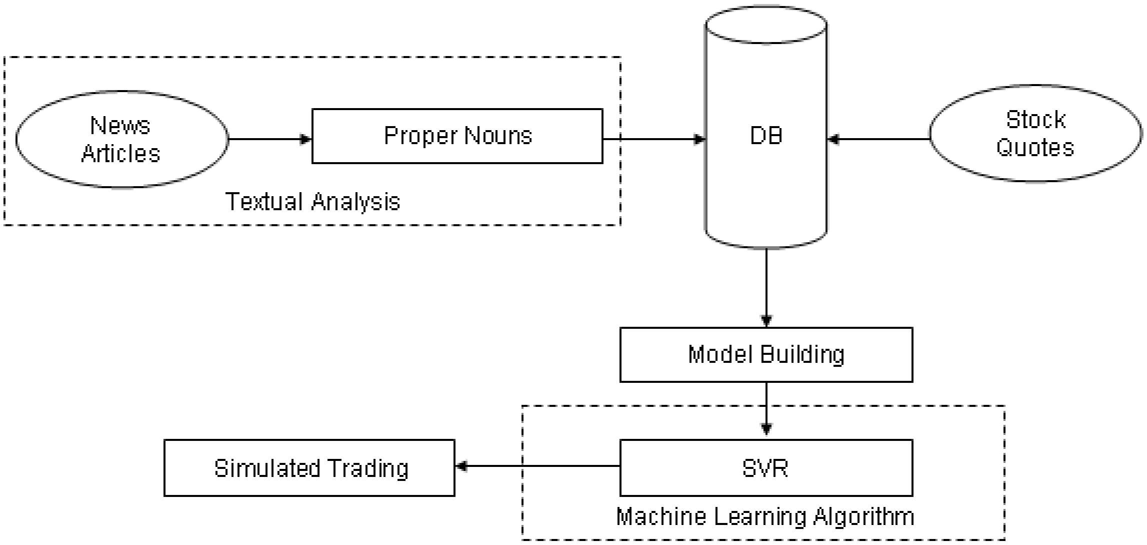


Figure 2. AZFinText system design

The differences in our design rest on several key factors; a Proper Noun representational scheme and an SVM regression derivative that outputs discrete numeric values. Financial news articles are gathered from Yahoo Finance and are represented by their Proper Nouns, such as specific people or organizations. To limit the size of the feature space, we only selected Proper Nouns that occur three or more times in a document.

Once the text has been represented and per minute stock quotation data has been obtained, the next step is to build a model for machine learning. In prior work at the University of Arizona, various models of data were tested with this goal in mind and from our results it was found that using both the article terms and the stock price at the time of article release led to the best predictive results [9]. After building the model, the data is used to train the machine learning algorithm and results are evaluated.

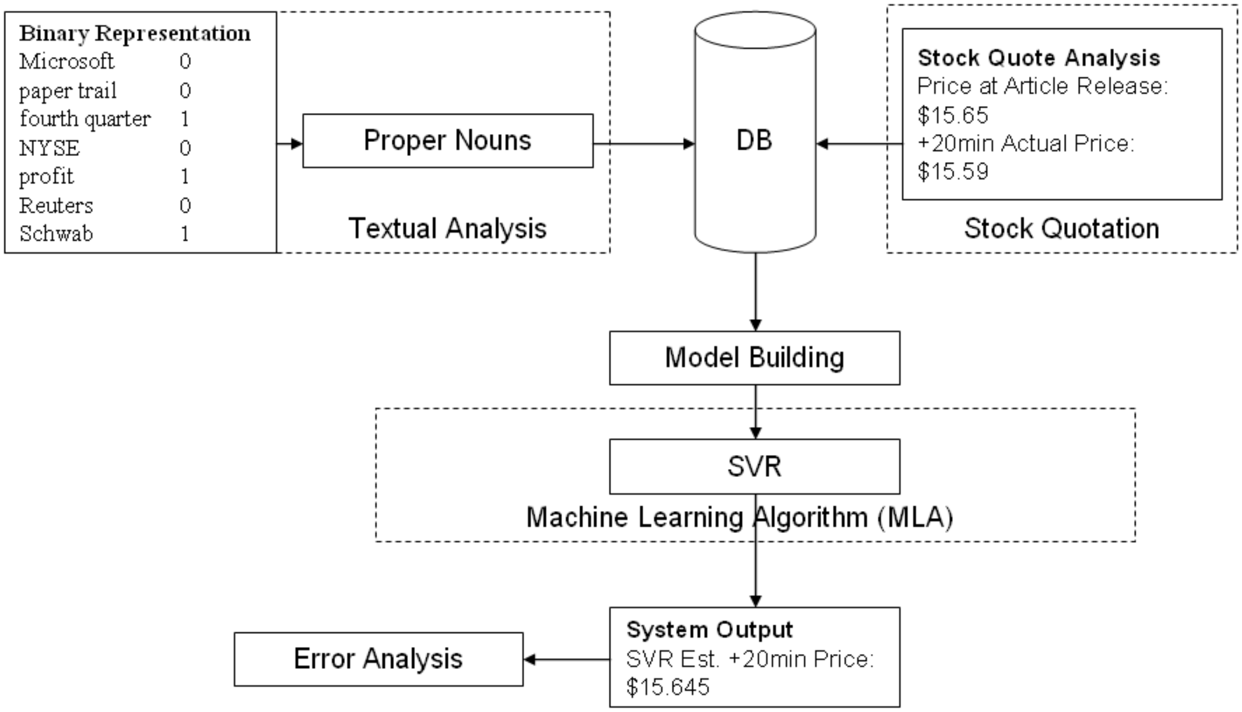


Figure 3. Example AZFinText representation

Figure 3 shows an example usage of the AZFinText system. The first step is to extract all article terms from every article in the corpora. Second, terms are identified by their parts of speech. Third, the entire set of Proper Nouns are represented in binary as either present or not in each individual article. Fourth, the price of the stock at the time the article was released, is then appended to each news article at the model building stage. For this particular article, the price of Schwab stock was $15.65 at the time of article release and the +20 minute stock price was $15.59. We selected the +20 minute interval to make our predictions because of its prior representation as a small window of opportunity in textual financial research [8, 9]. Once the model is built, machine learning takes place via the SVR algorithm. The SVR is fed a matrix of proper nouns, coded in binary as present or not in the article, as well as the price of the stock at the time the article was released. This is done for each textual financial news article and the SVR component makes a discrete prediction of what the +20 minute stock should be. In this instance, the output price is $15.645.

After training we analyze the data using a Simulated Trading Engine that invests $1,000 per trade and will buy/short the stock if the predicted +20 minute stock price is greater than or equal to 1% movement from the stock price at the time the article was released [4, 8, 9]. Any bought/shorted stocks are then sold after 20 minutes.

Research Testbed

For our experiment, we selected a consecutive five-week research period from Oct. 26, 2005 to Nov. 28, 2005 and limited the scope of activity to companies listed in the S&P 500 as of Oct. 3, 2005. This five-week period of study gathered 9,211 financial news articles, which is comparable to prior studies [8, 9]. Articles gathered during this period were further restricted to occur between the hours of 10:30am and 3:40pm. While trading starts at 9:30am, we felt it was important to reduce the impact of overnight news on stock prices and selected a period of one-hour to allow prices to adjust. The 3:40pm cut-off for news articles was selected to disallow any +20 minute stock predictions to occur after market hours. A further constraint was introduced to reduce the effects of confounding variables, where two articles on the same company cannot exist within twenty minutes of each other or both will be discarded. The above processes had reduced the 9,211 candidate news articles to 2,809, where the majority of discarded articles occurred outside of market hours. Similarly, 10,259,042 per-minute stock quotations of the same trading period were gathered from a commercial system.

In the model-building stage, financial news articles are aggregated together by their Industry Sectors prior to training. Financial trading analyses often use Sector-based comparisons in evaluating individual company performance. For AZFinText, articles are partitioned using the Global Industry Classification Standard (GICS) system of classification developed by Morgan Stanley. Companies from the S&P 500 are assigned an eight-digit GICS classifier which is used to identify sector, industry group, industry and sub-industry categories. Articles in each GICS Sector (10 Sectors in total) are then sent to AZFinText separately for 10-fold cross-validation. Each Sector averaged 281 articles with a standard deviation per category of 160.8.

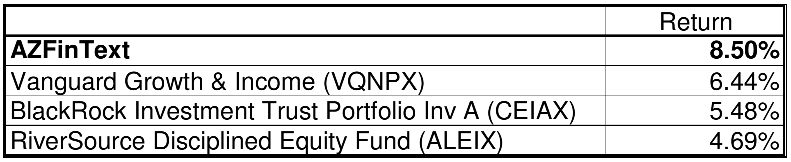
Testing AZFinText: Experimental Results

With some 90+ quant funds operating for a full year at the time of our study, we selected the top 10 quant funds according to their trailing one-year returns [2] to compare against AZFinText during our trading period. We also compared our AZFinText system against the S&P 500 index, which is the industry benchmark of performance. The results are presented in Table 1.

	Return
ProFunds Ultra Japan Inv (UJPIX)	24.73%
ProFunds Ultra Japan Svc (UJPSX)	24.59%
American Century Global Gold Adv (ACGGX)	12.96%
American Century Global Gold Inv (BGEIX)	12.93%
AZFinText	**8.50%**
Quantitative Advisors Emerging Markets Instl (QEMAX)	8.16%
Quantitative Advisors Emerging Markets Shs (QFFOX)	8.15%
S&P 500 Index	**5.62%**
Lord Abbett Small-Cap Value Y (LRSYX)	5.22%
Lord Abbett Small-Cap Value A (LRSCX)	5.19%
Quantitative Advisors Foreign Value Instl (QFVIX)	4.99%
Quantitative Advisors Foreign Value Shs (QFVOX)	4.95%

Table 1. Simulated Trading results of the Top 10 Quants

From Table 1, AZFinText had an 8.50% return on trades versus the S&P 500 of 5.62% during the same period. Comparing AZFinText against the top 10 quants shows AZFinText performing well, outperforming 6 of the top 10 quant funds. It is interesting to note that the four better performing quants were trading in the Nikkei and gold markets where AZFinText was constrained to the companies in the S&P 500. In making a more direct performance comparison, Table 2 shows the trade returns of AZFinText versus several quant funds that are also operating within the S&P 500.



	Return
AZFinText	**8.50%**
Vanguard Growth & Income (VQNPX)	6.44%
BlackRock Investment Trust Portfolio Inv A (CEIAX)	5.48%
RiverSource Disciplined Equity Fund (ALEIX)	4.69%

Table 2. Simulated Trading results of S&P 500 quants

As shown in this table, AZFinText performed better than its peer quant funds. It is worthwhile to point out that AZFinText's success came mostly from making predictions from financial news articles and stock quotes, whereas quants used sophisticated mathematical models on a large set of financial variables. We believe that our research helps identify a promising research direction in financial text mining. However, more research is critically needed.

Conclusion

The performance of AZFinText against existing quant funds is surprisingly robust given AZFinText's indifference to the internal fiscal makeup of the companies traded. We believe that this approach may encourage quant traders to incorporate a financial news analysis engine into existing strategies. Future quant funds can be more flexible and potentially more robust by obviating risks that are captured by unexpected news events. Future directions for this research include relaxing certain assumptions and carefully testing their impact on prediction as well as testing newer machine learning techniques such as probabilistic modeling (e.g., gaussian process) and adaptive boosting classifiers (e.g., adaboost).

References

1. Brachman, R.J., Khabaza, T., Kloesgen, W., Piatetsky-Shapiro, G. and Simoudis, E. Mining Business Data. *Communications of the ACM*, *39* (11).
2. Burke, K. Not the Man, But the Machine, http://www.registeredrep.com/moneymanagers/finance_not_man_machine/index.html, 2006.
3. Cai, G. and Wurman, P. Monte Carlo Approximation in Incomplete Information, Sequential Auction Games. *Decision Support Systems*, *39* (2). 153-168.
4. Fung, G.P.C., Yu, J.X., Yu, X. and Lam, W., News Sensitive Stock Trend Prediction. in *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, (Taipei, Taiwan, 2002).
5. Jelveh, Z. How a Computer Knows What Many Managers Don't *The New York Times*, 2006.
6. LeBaron, B., Arthur, W.B. and Palmer, R. Time Series Properties of an Artificial Stock Market. *Journal of Economic Dynamics and Control*, *23* (9-10). 1487-1516.
7. Lucchetti, A. and Lahart, J. Your Portfolio on AutoPilot; Brokerages Roll Out Software to Automate Trading Strategies; Risks of Becoming a 'Quant' *Wall Street Journal*, New York, 2006, B1.
8. Mittermayer, M., Forecasting Intraday Stock Price Trends with Text Mining Techniques. in *Proceedings of the 37th Hawaii International Conference on System Sciences*, (Hawaii, 2004).
9. Schumaker, R.P. and Chen, H., Textual Analysis of Stock Market Prediction Using Financial News Articles. in *12th Americas Conference on Information Systems (AMCIS-2006)*, (Acapulco, Mexico, 2006).
10. Schumaker, R.P. and Chen, H., Textual Analysis of Stock Market Prediction Using Financial News Articles. in *Americas Conference on Information Systems*, (Acapulco, Mexico, 2006).
11. Stauffer, D. EconoPhysics - A New Area for Computational Statistical Physics? *International Journal of Modern Physics C*, *11* (6). 1081-1087.
12. Tay, F. and Cao, L. Application of Support Vector Machines in Financial Time Series Forecasting. *Omega*, *29*. 309-317.
13. Vapnik, V. *The Nature of Statistical Learning Theory*. Springer, New York, 1995.
14. Yoon, Y. and Swales, G., Predicting Stock Price Performance: A Neural Network Approach. in *Proceedings of the 24th Hawaii International Conference on System Sciences*, (Hawaii, 1991), 156-162 vol.154.