The International Arab Conference on Information Technology (ACIT'2013)

PREDICTING STOCK PRICES USING DATA MINING TECHNIQUES

[1] QASEM A. AL-RADAIDEH, [2] ADEL ABU ASSAF [3] EMAN ALNAGI

[1]Department of Computer Information Systems, Faculty of Information Technology and Computer Science
Yarmouk University, Irbid, Jordan. {qasemr@yu.edu.jo}

[2]ICT Department, Amman Stock Exchange, Amman, Jordan. {abuassaf@gmail.com}

[3]Department of Computer Science, Faculty of Information Technology
Philadelphia University, Jordan {ealnagi@philadelphia.edu.jo}

ABSTRACT

Forecasting stock return is an important financial subject that has attracted researchers' attention for many years. It involves an assumption that fundamental information publicly available in the past has some predictive relationships to the future stock returns. This study tries to help the investors in the stock market to decide the better timing for buying or selling stocks based on the knowledge extracted from the historical prices of such stocks. The decision taken will be based on decision tree classifier which is one of the data mining techniques. To build the proposed model, the CRISP-DM methodology is used over real historical data of three major companies listed in Amman Stock Exchange (ASE).

Keywords: *Data Mining, Data Mining, Data Classification, Decision Tree, Future stock return, data mining techniques, decision tree classifiers, CRISP-DM methodology, Amman Stock Exchange.*

1. INTRODUCTION

The stock market is essentially a non-linear, non-parametric system that is extremely hard to model with any reasonable accuracy [1]. Investors have been trying to find a way to predict stock prices and to find the right stocks and right timing to buy or sell. To achieve those objectives, and according to [2], [3-4] some research used the techniques of fundamental analysis, where trading rules are developed based on the information associated with macroeconomics, industry, and company. The authors of [5] and [6] said that fundamental analysis assumes that the price of a stock depends on its intrinsic value and expected return on investment. Analyzing the company's operations and the market in which the company is operating can do this. Consequently, the stock price can be predicted reasonably well. Most people believe that fundamental analysis is a good method only on a long-term basis. However, for short- and medium-term speculations, fundamental analysis is generally not suitable.

Some other research used the techniques of technical analysis [2], in which trading rules were developed based on the historical data of stock trading price and volume. Technical analysis as illustrated in [5] and [7] refers to the various methods that aim to predict future price movements using past stock prices and volume information. It is based on the assumption that history repeats itself and that future market directions can be determined by examining historical price data. Thus, it is assumed that price trends and patterns exist that can be identified and utilized for profit. Most of the techniques used in technical analysis are highly subjective in nature and have been shown not to be statistically valid.

Recently, data mining techniques and artificial intelligence techniques like decision trees, rough set approach, and artificial neural networks have been applied to this area [8]. Data mining [9] refers to extracting or mining knowledge from large data stores or sets. Some of its functionalities are the discovery of concept or class descriptions, associations and correlations, classification, prediction, clustering, trend analysis, outlier and deviation analysis, and similarity analysis. Data classification can be done in many different methods; one of those methods is the classification by using Decision Tree. It is a graphical representation of all possible outcomes and the paths by which they may be reached.

Decision trees and artificial neural networks can be trained by using an appropriate learning algorithm.

Following the assumption of technical analysis that patterns exist in price data, it is possible in principle to use data mining techniques to discover these patterns in an automated manner. Once these patterns have been discovered, future prices can be predicted.

Today, the grand challenge of using a database is to generate useful rules from raw data in a database for users to make decisions, and these rules may be hidden deeply in the raw data of the database. Traditionally, the method of turning data into knowledge relies on manual analysis; this is becoming impractical in many domains as data volumes grow exponentially. The problem with predicting stock prices is that the volume of data is too large and huge. This paper uses one of the data mining methods; which is the classification approach on the historical data available to try to help the investors to build their decision on whether to buy or sell that stock in order to achieve profit.

The main objective of this paper is to analyze the historical data available on stocks using decision tree technique as one of the classification methods of data mining in order to help investors to know when to buy new stocks or to sell their stocks.

Analyzing stock price data over several years may involve a few hundreds or thousands of records, but these must be selected from millions. The data that will be used in this paper to build the decision tree will be the historical prices of three listed companies in Amman Stock Exchange over two years of time.

The remainder of this paper is organized into four sections. Section 2 of the paper gives a literature review about the subject of using data mining techniques in order to try to predict the prices and the trend of stocks, some related work in that subject is shown in this section. Section 3 talks about the methodology used in building the classification model. Then section 4 shows the experiments that are done on the data collected using the model and evaluation of the results using one of the evaluation methods. Finally, a brief conclusion and the future work about the topic is given in section 5.

2. LITERATURE REVIEW

Over the past two decades many important changes have taken place in the environment of financial markets. The development of powerful communication and trading facilities has enlarged the scope of selection for investors. Forecasting stock return is an important financial subject that has attracted researchers' attention for many years. It involves an assumption that fundamental information publicly available in the past has some predictive relationships to the future stock returns [10]. In order to be able to extract such relationships from the available data, data mining techniques are new techniques that can be used to extract the knowledge from this data.

For that reason, several researchers have focused on technical analysis and using advanced math and science. Extensive attention has been dedicated to the field of artificial intelligence and data mining techniques [11]. Some models have been proposed and implemented using the above mentioned techniques, the authors of [5] made an empirical study on building a stock buying/selling alert system using back propagation neural networks (BPNN), their NN was codenamed NN5. The system was trained and tested with past price data from Hong Kong and Shanghai Banking Corporation Holdings over the period from January 2004 to December 2005. The empirical results showed that the implemented system was able to predict short-term price movement directions with accuracy about 74%.

The research by [2] used decision tree technique to build on the work of Lin [12] where Lin tried to modify the filter rule that is to buy when the stock price rises k% above its past local low and sell when it falls k% from its past local high. The proposed modification to the filter rule in [12] was by combining three decision variables associated with fundamental analysis. An empirical test, using the stocks of electronics companies in Taiwan, showed Lin's method outperformed the filter rule. According to [2], in Lin's work, the criteria for clustering trading points involved only the past information; the future information was not considered at all. The research by [2] aimed to improve the filter rule and Lin's study by considering both the past and the future information in clustering the trading points. The researchers used the data of Taiwan stock market and that of NASDAQ to carry out empirical tests. Test results showed that the proposed method outperformed both Lin's method and the filter rule in the two stock markets.

The model of [11] applied the concept of serial topology and designed a new decision system, namely the two-layer bias decision tree, for stock price prediction. The methodology developed by the authors differs from other studies in two respects; first, to reduce the classification error, the decision model was modified into a bias decision model. Second, a two-layer bias decision tree is used to improve purchasing accuracy. The empirical results indicated that the presented decision model produced excellent purchasing accuracy, and it significantly outperformed than random purchase.

The authors of [10] presented an approach that used data mining methods and neural networks for forecasting stock market returns. An attempt has been made in this study to investigate the predictive power of financial and economic variables by adopting the variable relevance analysis technique in machine learning for data mining.

The authors examined the effectiveness of the neural network models used for level estimation and classification. The results showed that the trading strategies guided by the neural network classification models generate higher profits under the same risk exposure than those suggested by other strategies.

The research by [13] was basically a comparison between the work of Fama and French's model [14-15] and the artificial neural networks in order to try to predict the stock prices in the Chinese market. The purpose of this study is to demonstrate the accuracy of ANN in predicting stock price movement for firms traded on the Shanghai Stock Exchange. In order to demonstrate the accuracy of ANN, the authors made a comparative analysis between Fama and French's model and the predictive power of the univariate and multivariate neural network models. The results from this study indicated that artificial neural networks offer an opportunity for investors to improve their predictive power in selecting stocks, and more importantly, a simple univariate model appears to be more successful at predicting returns than a multivariate model.

Al-Haddad et al., [16] presented a study that aimed to provide evidence of whether or not the corporate governance & performance indicators of the Jordanian industrial companies listed at Amman Stock Exchange (ASE) are affected by variables that were proposed and to provide the important indicators of the relationship of corporate governance & firms' performance that can be used by the Jordanian industrial firms to solve the agency problem. The study random sample consists of (44) Jordanian industrial firms. The study founds a positive direct relationship between corporate governance and corporate performance.

Hajizadeh et al. [17] provided an overview of application of data mining techniques such as decision tree, neural network, association rules, and factor analysis and in stock markets.

Prediction stock price or financial markets has been one of the biggest challenges to the AI community. Various technical, fundamental, and statistical indicators have been proposed and used with varying results. Soni [18] surveyed some recent literature in the domain of machine learning techniques and artificial intelligence used to predict stock market movements. Artificial Neural Networks (ANNs) are identified to be the dominant machine learning technique in stock market prediction area.

El-Baky et al., [19], proposed a new approach for fast forecasting of stock market prices. The proposed approach uses new high speed time delay neural networks (HSTDNNs). The authors used the MATLAB tool to simulate results to confirm the theoretical computations of the approach.

3. THE METHODOLOGY OF THE STUDY

Data mining methodology is designed to ensure that the data mining effort leads to a stable model that successfully addresses the problem it is designed to solve. Various data mining methodologies have been proposed to serve as blueprints for how to organize the process of gathering data, analyzing data, disseminating results, implementing results, and monitoring improvements [9]. To build the model that analyses the stock trends using the decision tree technique, the CRISP-DM (Cross-Industry Standard Process for data mining) [20] is used. This methodology was proposed in the mid-1990s by an European consortium of companies to serve as a non-proprietary standard process model for data mining. This model consists of the following six steps:

- Understanding the reason and objective of mining the stock prices.
- Understanding the collected data and how it is structured.
- Preparing the data that is used in the classification model.
- Selecting the technique to build the model.
- Evaluating the model by using one of the well known evaluation methods.
- Deploying the model in the stock market to predict the best action to be taken, either selling or buying the stocks.
- Understanding the reason and objective of building the model

The main reason and objective of building the model is to try to help the investors in the stock market to decide the best timing for buying or selling stocks based on the knowledge extracted from the historical prices of such stocks. The decision taken will be based on one of the data mining techniques; the decision tree classifiers.

Understanding the collected data

The Oracle database of Amman Stock Exchange (ASE) contains the historical prices of the 230 companies listed in the exchange from the year 2000. As the amount of such data is very large and complicated, the decision was taken to choose three companies listed in the exchange. The selection of these companies was based on the following five criteria which represent the companies' size and liquidity: Market capitalization, days traded, turnover ratio, value traded and the number of shares traded, also the sector representation was considered during the selection of these companies. These companies are "Arab Bank", its' code in the stock market "ARBK" and it

belongs to the banking sector, "United Arab Investors Company", its' code is "UAIC" and it belongs to the services sector, and "Middle East Complex for Engineering, Electronics and Heavy Industries", its' code is "MECE" and it belongs to the industrial sector. The period that was selected is from April 2005 to May 2007, which presented the current and actual status of the market at that period of time.

At the beginning, the data collected contained 9 attributes; this number was reduced manually to 6 attributes as the other attributes were found not important and not having a direct effect on the study. Table1 shows the 6 attributes selected with their descriptions and their possible values. The class attribute is the investor action whether to buy or sell that stock and it is named, "Action". The data of this attribute was taken also from ASE database, which is the net position of one of the biggest brokers dealing with the above mentioned stocks every day. The net position could be either buying or selling that stock for that day.

Table 1: Attribute Description

Attribute	Description	Possible Values
Previous	Previous day close price of the stock	Positive, Negative, Equal
Open	Current day open price of the stock	Positive, Negative, Equal
Min	Current day minimum price of the stock	Positive, Negative, Equal
Max	Current day maximum price of the stock	Positive, Negative, Equal
Last	Current day close price of the stock	Positive, Negative, Equal
Action	The action taken by the investor on this stock	Buy, Sell

Preparing the data

At the beginning, when the data was collected, all the values of the attributes selected were continuous numeric values. Data transformation was applied by generalizing data to a higher-level concept so as all the values became discrete. The criterion that was made to transform the numeric values of each attribute to discrete values depended on the previous day closing price of the stock. If the values of the attributes open, min, max, last were greater than the value of attribute previous for the same trading day, the numeric values of the attributes were replaced by the value Positive. If the values of the attributes mentioned above were less than the value of the attribute previous, the numeric values of the attributes were replaced by Negative. If the values of those attributes were equal to the value of the attribute previous, the values were replaced by the value Equal. Table 2 shows a sample of the continuous numeric values of the data before selecting the 6 attributes manually and before generalizing them to discrete values, while table3 shows the same sample after selecting the 6 attributes and after transforming them to discrete values.

Building the model

After the data has been prepared and transformed, the next step was to build the classification model using the decision tree technique. The decision tree technique was selected because [9] the construction of decision tree classifiers does not require any domain knowledge, thus it is appropriate for exploratory knowledge discovery. Also, it can handle high dimensional data. Another benefit is that the steps of decision tree induction are simple and fast. Generally, decision tree accuracy is considered good. The decision tree method depends on using the information gain metric that determines the most useful attribute. The information gain depends on the entropy measure.

Table 2: Sample of historical data before selecting relevant attributes and before generalization

Previous	Open	Max	Min	Last	Action
25.82	25.99	26	25.41	25.67	Sell
25.67	25.68	25.68	25.2	25.3	Buy
25.3	24.8	25.3	24.41	24.9	Buy
24.9	24.8	24.9	24.3	24.87	Sell
24.87	24.87	25.55	24.85	25.3	Buy
25.3	25.25	26	25.25	25.82	Buy
25.82	25.99	26.4	25.99	26.3	Buy

26.3	26.3	26.7	26	26.02	Buy
26.02	26.09	26.25	25.55	25.63	Sell

Table 3: Sample of historical data after selecting attributes and after generalization.

Previous	Open	Max	Min	Last	Action
Positive	Positive	Positive	Negative	Negative	**Sell**
Negative	Positive	Positive	Negative	Negative	**Buy**
Negative	Negative	equal	Negative	Negative	**Buy**
Negative	Negative	equal	Negative	Negative	**Sell**
Negative	equal	Positive	Negative	Positive	**Buy**
Positive	Negative	Positive	Negative	Positive	**Buy**
Positive	Positive	Positive	Positive	Positive	**Buy**
Positive	equal	Positive	Negative	Negative	**Buy**
Negative	Positive	Positive	Negative	Negative	**Sell**

The gain ratio is used to rank attributes and to build the decision tree where each attribute is located according to its gain ratio. When the decision tree model was applied on the data of the three companies using the WEKA software version 3.5 [21], the root attribute for both ARBK and UAIC company was the Open, while the attribute Last was the root for the decision tree of the MECE company. As the process of building the tree goes on, all the remaining attributes were used to continue with this process. After building the complete decision tree, the set of classification rules were generated by following all the paths of the tree. The maximum number of attributes that were used in some of the classification rules generated were 4 attributes, while some classification rules used only 1 attribute. Both the ID3 and C4.5 algorithms were used in building the decision trees and the pruning technique was used in the C4.5 algorithm in order to reduce the size of the produced decision trees. Table 4 gives a summary about the numbers of the classification rules that resulted after building the decision trees for each company using the C4.5 algorithm.

The graphs of the resulting decision trees using the C4.5 algorithm with pruning technique is presented in Figure 1, Figure 2, and Figure 3 for the three companies under study.

Table 4: Summary of the number of the classification rules

Company	Number of classification rules without pruning	Number of classification rules with pruning
ARBK	21	11
UAIC	31	5
MECE	21	9

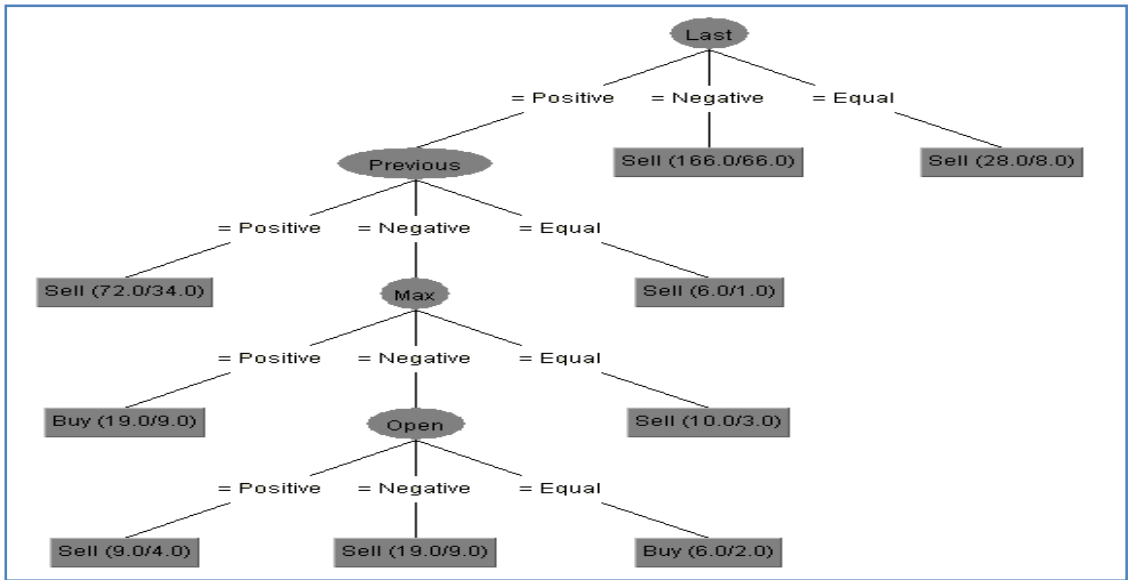

Figure 1: Decision Tree for the MECE

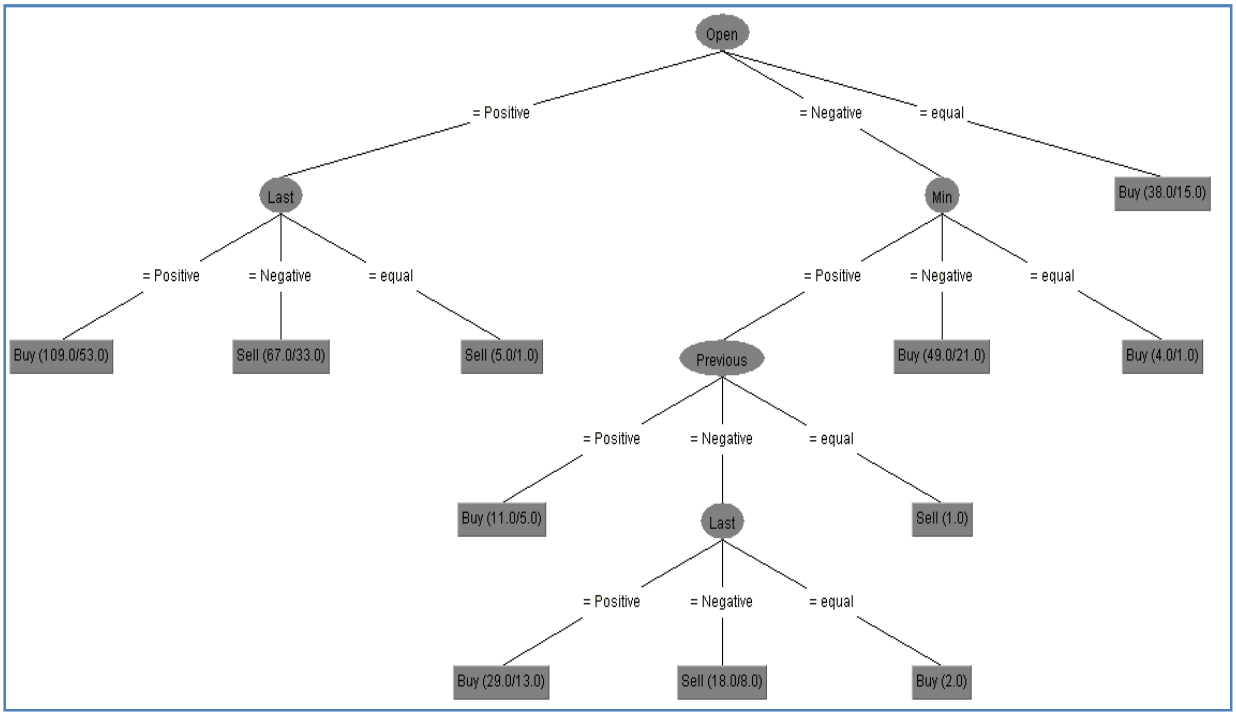

Figure 2: The Decision Tree for the ARBK

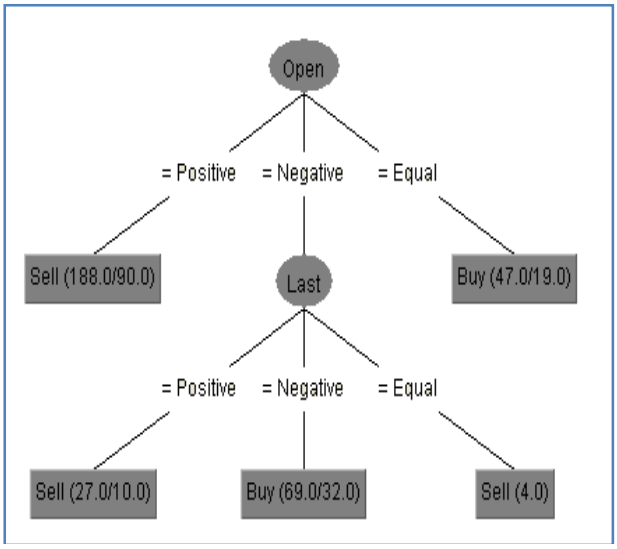

Figure 3: The Decision Tree for the UAIC

Deploying the model

The classification rules that were generated from the decision tree model can be used and integrated in a system that predict the best action and timing for the investors, either to buy or sell the stocks on that day.

4. RESULTS AND DISCUSSION

This section presents step 5 of the CRISP methodology which was used to build the model. It is simply about evaluating the model by using one or more of the well known evaluation methods. In order to evaluate the model, the WEKA software was used to calculate the accuracy of the classification model. Two evaluation methods were used, the K-Fold Cross Validation (K-CV) where K=10 folds and the percentage split method where 66% of the data was used for training and the remainder for testing. Both evaluation methods were used on the ID3 and C4.5 decision tree classification methods. Table 5 shows the accuracy of all the classifiers generated using both classification methods and both evaluation methods.

As we can see from the table, the resultant classification accuracy from the decision tree model is not very high for the training data used and it varies from one company to another. The reason for such a low accuracy is that the company's performance in the stock market is affected by internal financial factors such as; news about the company, financial reports, and the overall performance of the market. Also, external factors can affect the performance of the company in the market such as; political events and political decisions. Thus, it can be difficult to have a model that gives a high accuracy classification for all the companies at the same time as the performance of these companies differs.

Table 5: Classification accuracy using ID3 & C4.5 classification methods and using 10-CV& Holdout evaluation methods

		10-CV			Holdout 66%		
Company	**Classification Method**	**Total Instances**	**Correctly classified**	**Accuracy %**	**Total Instances**	**Correctly classified**	**Accuracy %**
ARBK	**ID3**	499	233	44.689	170	73	42.941
	C4.5		237	47.495		83	48.824
MECE	**ID3**	502	255	50.797	171	84	49.123
	C4.5		265	52.789		91	53.216
UAIC	**ID3**	502	269	53.586	171	88	51.462
	C4.5		264	52.590		94	54.971

5. CONCLUSIONS AND FUTURE WORK

This study presents a proposal to use the decision tree classifier on the historical prices of the stocks to create decision rules that give buy or sell recommendations in the stock market. Such proposed model can be a helpful tool for the investors to take the right decision regarding their stocks based on the analysis of the historical prices of stocks in order to extract any predictive information from that historical data. The results for the proposed model were not perfect because many factors including but not limited to political events, general economic conditions, and investors' expectations influence stock market.

As for the future work, there is still big room for testing and improving the proposed model by evaluating the model over the whole companies listed in the stock market. Also, the evaluation of a larger collection of learning techniques such as neural networks, genetic algorithms, and association rules can represent a rich area for future investigation. Finally, reconsidering the factors affecting the behavior of the stock markets, such as trading volume, news and financial reports which might impact stock price can be another rich field for future studying.

REFERENCES

[1] Wang, Y.F., (2003) "Mining stock price using fuzzy rough set system", Expert Systems with Applications, 24, pp. 13-23.

[2] Wu, M.C., Lin, S.Y., and Lin, C.H., (2006) "An effective application of decision tree to stock trading", Expert Systems with Applications, 31, pp. 270-274.

[3] Al-Debie, M., Walker, M. (1999). "Fundamental information analysis: An extension and UK evidence", Journal of Accounting Research, 31(3), pp. 261–280.

[4] Lev, B., Thiagarajan, R. (1993). "Fundamental information analysis", Journal of Accounting Research, 31(2), 190–215.

[5] Tsang, P.M., Kwok, P., Choy, S.O., Kwan, R., Ng, S.C., Mak, J., Tsang, J., Koong, K., and Wong, T.L. (2007) "Design and implementation of NN5 for Hong Kong stock price forecasting", Engineering Applications of Artificial Intelligence, 20, pp. 453-461.

[6] Ritchie, J.C., (1996) Fundamental Analysis: a Back-to-the-Basics Investment Guide to Selecting Quality Stocks. Irwin Professional Publishing.

[7] Murphy, J.J., (1999) Technical Analysis of the Financial Markets: a Comprehensive Guide to Trading Methods and Applications. New York Institute of Finance.

[8] Wang, Y.F., (2002) "Predicting stock price using fuzzy grey prediction system", Expert Systems with Applications, 22, pp. 33-39.

[9] Han, J., Kamber, M., Jian P. (2011). "Data Mining Concepts and Techniques". San Francisco, CA: Morgan Kaufmann Publishers.

[10] Enke, D., Thawornwong, S. (2005) "The use of data mining and neural networks for forecasting stock market returns", Expert Systems with Applications, 29, pp. 927-940.

[11] Wang, J.L., Chan, S.H. (2006) "Stock market trading rule discovery using two-layer bias decision tree", Expert Systems with Applications, 30(4), pp. 605-611.

[12] Lin, C. H. (2004) Profitability of a filter trading rule on the Taiwan stock exchange market. Master thesis, Department of Industrial Engineering and Management, National Chiao Tung University.

[13] Cao, Q., Leggio, K.B., and Schniederjans, M.J., (2005) "A comparison between Fama and French's model and artificial neural networks in predicting the Chinese stock market", Computers & Operations Research, 32, pp. 2499-2512.

[14] Fama, E.F., French, K.R., (1993) "Common risk factors in the returns on stocks and bonds", The Journal of Finance, 33, pp. 3-56.

[15] Fama, E.F., French, K.R., (1992) "The cross-section of expected stock returns", The Journal of Finance, 47, pp. 427-465.

[16] Al-Haddad W. Alzurqan S. and Al_Sufy S, The Effect of Corporate Governance on the Performance of Jordanian Industrial Companies: An empirical study on Amman Stock Exchange. International Journal of Humanities and Social Science, Vol. 1 No. 4; April 2011

[17] Hajizadeh E., Ardakani H., and Shahrabi J., Application of data mining techniques in stock markets: A survey, Journal of Economics and International Finance Vol. 2(7), pp. 109-118, July 2010.

[18] Soni S., Applications of ANNs in Stock Market Prediction: A Survey, International Journal of Computer Science & Engineering Technology (IJCSET), pp 71-83, Vol. 2 No. 3, 2011.

[19] Hazem M. El-Bakry, and Wael A. Awad, Fast Forecasting of Stock Market Prices by using New High Speed Time Delay Neural Networks, International Journal of Computer and Information Engineering 4:2 2010. Pp 138-144.

[20] Chapman P., Clinton J., Kerber R., Khabaza T., Reinartz T., Shearer C., and Wirth R., (2000). "CRISP-DM 1.0: Step-by-step data mining guide".

[21] Witten I. Frank E., and Hall M. (2011), "Data Mining: Practical Machine Learning Tools and Techniques", 3rd Edition, Morgan Kaufmann Publishers.