International Academic Institute for Science and Technology

International Academic Journal of Economics
Vol. 2, No. 2, 2015, pp. 1-11.

ISSN 2454-2474

International Academic Journal of Economics

www.iaiest.com

Predict the trend of stock prices using machine learning techniques

[a]Seyed Enayatolah Alavi, [b]Hasanali Sinaei, [c]Elham Afsharirad

[a]Assistant Profossor of Shahid Chamran universty of Ahvaz, Ahvaz, Iran.

[b]*Associate Professor of Shahid Chamran universty of Ahvaz, Ahvaz, Iran.*

[c]*Master student Of Shahid Chamran universty of Ahvaz, Iran.*

Abstract

According to growing importance of the stock market in economic conditions of each country and since, stock prices are the most important factors influencing investment decisions for selecting stock, predicting the stock price movement is an integral part of the investment. This paper presented to forecast the movement of stock prices Tejarat bank of Iran with considerable precision. Accordingly, to predict the trend of stock prices using machine learning techniques and economic indicators have been considered. About 18,000 different indicators are presented, both simple moving average (SMA), weighted moving average (WMA), relative strength *indicator* (RSI) and moving average convergence divergence (MACD) indicator those are widely used in the stock market of Iran, have been chosen. The output of those is input three classifier, support vector machines, random forests and k-nearest neighbour. The outputs of the three classifier will be compared with each other. The results in this paper show that, respectively, random forest classifier, support vector machine and the k- nearest neighbours have the best accuracy in categories.

Keywords: technical indicators, support vector machine, random forest and k-nearest neighbour

Introduction:

Due to a lot of unknown information on the trend of stock price movement, it is difficult to predict. There are two types of analysis which investors perform before investing in a stock, first is the fundamental analysis, In this, investors looks at intrinsic value of stocks, performance of the industry, economy, political climate etc. to decide whether to invest or not. On the other hand, technical analysis is the evaluation of stocks by means of studying statistics generated by market activity, such as past prices and volumes. Technical analysts do not attempt to measure a security' s intrinsic value but instead it, use stock charts to identify patterns and trends that may suggest how a stock will behave in the future, Patel and shah (2015).

In the beginning of the 21st century however, some economists indicated that future stock prices are at least partially predictable. Therefore a lot of prediction algorithms have been explored and showed that stock price behaviour can indeed be predicted, Malkiel (2003).

Theoretical basis of research:

Enke and Thawornwong (2005), Kim et al (2003) and Wang and Chan (2006) states that, since years, many techniques have been developed to predict stock trends. Initially classical regression methods were used to predict stock trends. Since stock data can be categorized as no stationary time series data, non-linear machine learning techniques have also been used. Artificial Neural Networks (ANN) and Support Vector Machine (SVM) are two machine learning algorithms which are most widely used for predicting stock and stock price index movement. Each algorithm has its own way to learn patterns. ANN emulates functioning of our brain to learn by creating network of neurons.

Donaldson and Kamstra (1996), Refenes et al (1994) and Yoon et al (1993) states that, recent studies in the aforementioned area focus on the use of machine learning techniques, such as neural networks and decision trees, to build a prediction model for stock returns prediction. Among these studies, the neural networks technique has experienced the most consideration. Related works have shown that neural networks outperform many other statistical based techniques such as regression and discriminated analysis.

Research Background

Ou et al In 2010 for first time 4 machine learning technics, support vector machine, logistic regression, neural network and k nearest neighbours (KNN) used for predicate stock price.

Tsai et al, in the machine learning and pattern recognition classifier combination shows better performance than the single individual categories. Tsai in 2011 investigated the prediction performance that utilizes the classifier ensembles method to analyse stock returns. The hybrid methods of majority voting and bagging were considered. Moreover, performance using two types of classifier ensembles was compared with those using single baseline classifiers (i.e. neural networks, decision trees and logistic regression). The results indicated that multiple classifiers outperform single classifiers in terms of prediction accuracy and returns on investment.

In the Hafezia et al article, a new intelligent model in a multi-agent framework called bat-neural network multi-agent system (BNNMAS) to predict stock price. The model performs in a four layer multi-agent framework to predict eight years of DAX stock price in quarterly periods. The results show that BNNMAS significantly performs accurate and reliable.

Amin et al, have a one of study about prediction of stock index Tehran used neural network and genetic algorithm neural network that get input from nine indicator and finally showed that the model use hybrid of two algorithm neural network and genetic algorithm neural network has efficiently currency.

This paper aims to predict trends Iranian Tejarat bank stocks with high accuracy using useful indicators that have been consistent with the market index that results from them as input to the three classifiers support vector machines, random forests and k-nearest neighbours data and the output of the three algorithms to select the best predictor is compared.

Research methodology

The population, sample and sampling

Ten years of data of total stock price indices Tejarat bank of Iran from 2002 to 2012 is used in this study. All the data is obtained from http://www.tsetmc.com/ website. Close price of each day our entire data set.

This study uses 20% of the entire data as the parameter selection data. This data is used to determine design parameters of predictor models. Parameter selection data set is constructed by taking equal proportion of data from each of the ten years. The proportion of percentage wise increase and decrease cases in each year is also maintained. This sampling method enables parameter setting data set to be better representative of the entire data set. This parameter selection data is further divided into training and hold-out set. Each of the set consists of 10% of the entire data.

Research tools

Optimum parameters for predictor models are obtained by means of experiments on parameter selection data. Then, to compare the performance and accuracy the optimum parameters send to support vector machines, random forests and k-nearest neighbours. There are some technical indicators through which one can predict the future movement of stocks.

Four technical indicators calculated based on the formula as discussed in the Table 1 are given as inputs to predictor models. The values of all technical indicators are normalized in the range between [-1, +1].

Technical indicator

First two technical indicators are moving averages. The moving average (MA) is simple technical analysis tool that smooth out price data by creating a constantly updated average price Because of high accepted yield moving average indicators for two of these indicators were used in this project. Moving

average are simple technical analysis tools that create a graph of the average of consecutive price data of a pattern to answer investor. In this paper, to predict the short-term moving average ten-day simple moving average and exponential moving average is used.

As a general guideline, if the price is above the moving average then the trend is up. If the price is below a moving average the trend is down.

If current price is above the moving average values then the trend is 'up' and represented as '+1', and if current price is below the moving average values then the trend is 'down' and represented as '_1' [1].
RSI is generally used for identifying the overbought and oversold points. It ranges between 0 and 100. If the value of RSI exceeds 70 level, it means that the stock is overbought, so, it may go down in near future (indicating opinion '_1') and if the value of RSI goes below 30 level, it means that the stock is oversold, so, it may go up in near future (indicating opinion '+1') [1].

MACD follows the trend of the stock, i.e. if MACD goes up then stock price also goes up and vice-a-versa. So, if the value of MACD at time 't' is greater than the value at time 't_1', opinion on trend is 'up' and represented as '+1' and if the value of MACD at time 't' is less than value at time 't_1', opinion on trend is 'down' and represented as '_1' [1].

Table 2 shows summary statistics for the selected indicators. In this paper, we use Metastock software for technical analysis.

Table 1: Selected technical indicators & their formulas

Name	Formulas
SMA	$\frac{c_1 + c_2 + \cdots + c_n}{n}$
WMA	$\frac{c_1 + 2 * c_2 + \cdots + n * c_n}{1 + 2 + \cdots + n}$
RSI	$100\text{-}\frac{100}{1+(\sum_{i=0}^{n-1} up_{t-i}/n)/(\sum_{i=0}^{n-1} dw_{t-i}/n)}$
MACD	$\text{-}MACD(n)_{t-1})MACD(n)_{t-1} + \frac{2}{n+1} * (DIFF_t$

Table 2: shows summary statistics for the selected indicators

Indicator	Max	Min	Mean	Standard deviation
SMA	6217.37	935.38	3789.68	1047.47
WMA	6214.38	940.35	3789.69	1051.11
RSI	100.00	1.42	56.40	46.28
MACD	277.17	-357.33	13.52	-13.39

Machine Learning Algorithms

Support Vector Machines

The support vector machines (SVMs) were proposed by Vapnik in 1999. There are two main categories for support vector machines: support vector classification (SVC) and support vector regression (SVR). SVM is a learning system using a high dimensional feature space. Khemchandani and Chandra stated that in SVM, points are classified by means of assigning them to one of two disjoint half spaces, either in the pattern space or in a higher-dimensional feature space. Khemchandani et al (2009). The main objective of support vector machine is to identify maximum margin hyper plane. The idea is that the margin of separation between positive and negative examples is, Xu et al (2009).

It finds maximum margin hyper plane as the final decision boundary. Assume that $x_i \in R^d$; i = 1; 2; . . . ;N forms a set of input vectors with corresponding class labels $y_i \in \{+1, -1\}$; i = 1; 2; . . . ;N. SVM can map the input vectors $x_i \in R^d$ into a high dimensional feature space $\phi_{(x_i)} \in$ H. A kernel function $k(\boldsymbol{x_i}, \boldsymbol{x_j})$, performs the mapping. The resulting decision boundary is defined in equation 1.

$$f(x) = \text{sgn}(\sum_{i=1}^{N} y_i \alpha_i \, . \, k(\boldsymbol{x}, x_i) + b \quad (1)$$

Quadratic programming problem shown in equation 2–4 is solved to get the values of α_i.

Maximize:

$$\sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j . y_i y_j . k(x_i, x_j) \quad (2)$$

Subject to $0 \leq \alpha_i \leq c$ (3)

$$\sum_{i=1}^{N} \alpha_i y_i = 0 \qquad i = 1,2, \ldots 3 \quad (4)$$

The trade-off between margin and misclassification error is controlled by the regularization parameter c. The polynomial and radial basis kernel functions are used by us and they are shown in equation 5 and 6 respectively.

Polynomial Function: $k(x_i, x_j) = (x_i x_j + 1)^d$ (5)

Radial Basis Function:

$$k(x_i, x_j) = \exp(-\gamma||x_i - x_j||^2) \quad (6)$$

Where d is the degree of polynomial function and γ is the constant of radial basis function.

Random forest

Decision tree learning is one of the most popular techniques for classification. Its classification accuracy is comparable with other classification methods, and it is very efficient.
ID3 presented by Quinlan (1986), C4.5 presented by Quinlan (1993) and CART presented by Breiman et al (1984) are decision tree learning algorithms. Details can be found in article of Han et al (2006). Random forest belongs to the category of ensemble learning algorithms. It uses decision tree as the base learner of the ensemble.

The idea of ensemble learning is that a single classifier is not sufficient for determining class of test data. Reason being, based on sample data, classifier is not able to distinguish between noise and pattern. So it performs sampling with replacement such that given n trees to be learnt are based on these data set samples.

Also in our experiments, each tree is learnt using 3 features selected randomly. After creation of n trees, when testing data is used, the decision which majority of trees comes up with is considered as the final output. This also avoids problem of over-fitting. Our implementation of random forest algorithm is summarized in the Algorithm 1.

Algorithm 1- Implementation of random forest

Input: training set D, number of trees in the ensemble k **Output**: a composite model M*
1: for i = 1 to k do **2**: Create bootstrap sample Di by sampling D with replacement. **3**: Select 3 features randomly. **4**: Use Di and randomly selected three features to derive tree Mi. **5**: end for **6**: return M*.

Number of trees in the ensemble n is considered as the parameter of random forest. To determine it efficiently, it is varied from 10 to 200 with increment of 10 each time during the parameter setting experiments. For one stock, these settings of parameter yield a total of 20 treatments. The top three parameter values that resulted in the best average of training and holdout performances are selected as the top three random forest models for the comparison experiments.

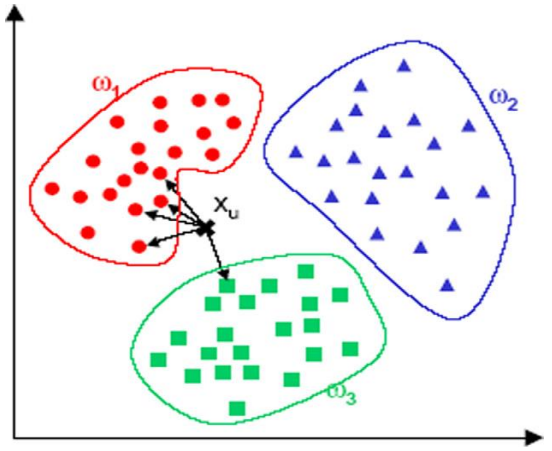
K-nearest neighbor

The K-nearest neighbor (KNN) classification approach is an instant-based learning algorithm that uses the nearest distance in determining the category of new vector in the training data set. During the training stage, the feature space is divided into multiple regions and the training data points are mapped into these regions according to the similarity of their contents. The unlabeled input data points are categorized to a particular category by finding the closet or distance from input data point and that particular category.

The KNN approach needs only a small number of training data points and this has contributed to the simplicity of the KNN which makes it outperforms other classification approaches, Han et al (1999).

Figure 1 shows an example of a 5-NN classifier which consists of three categories ω_1, ω_2 and ω_3. x_u is the new unlabeled input data point to be classified in the testing stage. The most commonly and widely used distance function for the KNN classifier is the Euclidean distance formula and it is used to calculate the distance between the new unlabeled data point and the training data points. The main step in the classification stage of the KNN is to measure the distance in order to identify the nearest neighbours of the new input data point, Han et al (1999).

Figure (1) Feature space of a 3-dimensional 5-NN classifier



According to figure 1, the value of parameter K is 5 and Euclidean distance formula has been used to calculate the distance between the training data points and the testing data point x_u. Among the five nearest neighbours of x_u, four are belong to category ω_1 and another one belongs to category ω_3. Hence, x_u is classified as category $\omega_1$by the KNN classifier.

The main advantage of the KNN is that it can perform well in classification tasks with multi categorized data points. On the other hand, since the KNN uses the distance calculation in determining the category of new input data points, this has brought to a great disadvantage of the KNN when the size of training set is big, where the classification model will become computationally intensive. This problem has made the KNN demands for more memory and CPU usages and high time consumption, especially in the classification stage. Moreover, this problem has also led to the drastic decrease in accuracy when there exist irrelevant features or noises in the training dataset. As an instant-based learning approach, the KNN freezes the training process until it receives a new input data point to be classified. The KNN compiles the entire training data points again when there is a new input sample and it discards the immediate result.

To counter the major drawback of the KNN approach, this is high time consumption, several enhancement techniques such as bucketing algorithm and k-dimensional trees algorithm have been introduced to improve the performance of the conventional KNN, Osuna (2002).

In this study, experiments have been done with K values and the number of different categories, the purpose of experiments on comparison data set is to compare the prediction performance of these models for best parameter combinations reported during parameter setting experiments.

During this comparison experiment, each of the prediction models is learnt based on best parameters reported by parameter setting experiments. Figure 2 depicts the prediction process.

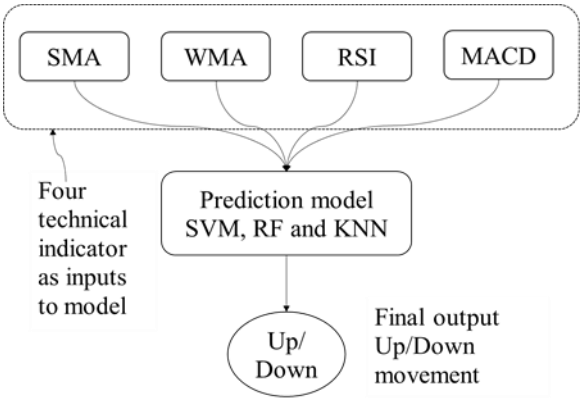


Figure (2) A schema of hole of prediction process

We've used MATLAB to execute learning machine techniques.

The Analysis of Data

We have used four economic indicator for first step of our analyzing data and three learning machine technics for second step of our analyzing data.

Research Findings

Accuracy and F-measure are used to evaluate the performance of proposed models. Computation of these evaluation measures requires estimating Precision and Recall which are evaluated from True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN). These parameters are defined in Equations (10) – (13).

$$Precision_{positive} = \frac{TP}{TP + FP} \quad (10)$$

$$Precision_{negative} = \frac{TP}{TP + FP} \quad (11)$$

$$Recall_{positive} = \frac{TP}{TP + FN} \quad (12)$$

$$Recall_{negative} = \frac{TN}{TN + FP} \quad (13)$$

Accuracy and F-measure are estimated using Equations (14) and (15) respectively.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (14)$$

$$\text{F-measure} = \frac{2 * precision * recall}{precision + recall} \quad (15)$$

The best parameter combinations are identified by means of experiments on parameter setting data set for each of the prediction models. These parameter combinations with corresponding accuracy and F-measure during parameter setting experiments are reported in tables 3-5.

Table 3-Best two parameter combinations (one for each type of kernel) of SVM model and their performance

	Kernel:Polynomial	Kernel:RBF
	c:100,degree:1	c:0.5,gamma:5
Accuracy	0.8427	0.8057
F-measure	0.8600	0.8275

Table 4- Best three parameter combinations of random forest model and their performance

	N_trees		
	140	20	30
Accuracy	0.9148	0.9146	0.9099
F-measure	0.9186	0.9185	0.9162

Table 5- Best three parameter combinations of KNN model and their performance

	K		
	25	45	70
Accuracy	0.8138	0.8059	0.8132
F-measure	0.8202	0.8135	0.8190

Tables show that the best average accuracy and F-measure get from Random forest and so from SVN and finally from KNN. It's because Random forest is an ensemble learning method for classification but

SVM and KNN are single classifier. Random decision forests correct for decision trees' habit of overfitting to their training set, Hastie et al (2009). It's relatively robust to outliers and noise. It gives useful internal estimates of: error, strength, correlation and variable importance.

Conclusion:

Stock market data is an example of non-stationary data. At particular time there can be trends, cycles, random walks or combinations of the three. In this study, for the first time in the stock market of the four

economic indicators to predict the stock price is three machine learning algorithm was used an Iranian company.

Because this study was not similar to the stock market's been done, the results compared with similar those are foreign. In an article of Ballings et al (2015), classifiers single neural network, support vector machines, logistic regression and k-nearest neighbour classifiers, Random Forest, and kernel factory are compared, the result of research is that a better outcome compared to single classifier ensemble random forest classifier, and between ensemble classifier also have the best results. This result is consistent with the results obtained in our research.

In an article for the first time a work of ten economic indicators was used is. Technical indicators has been ten first economic input, then the output of the ten indicators support vector machine regression and finally exits to the three neural network algorithm, random forests and support vector machine regression is given. This is as a result of the proposed model compared to the case where the input data to classifiers of data is used, the better. This result is consistent with the results obtained in our research.

Practical suggestions:

Because the stock market is influenced by many factors, political, economic, rumours, etc. These factors will also be in the future to predict the trend of stock prices of various companies, especially in the Tejarat Bank of Iran.

References:

Amin, Vahid, Salehnezhad S. Hasan, Valipour Mehrdad and Nasirlu Saber. (2014). Predicting Direction of Stock Price Index Volatility Using Genetic Algorithms and Artificial Neural Network Models in Tehran Stock Exchange. International Journal of Business and Technopreneurship. Vol. 4. No. 3. 451-465.

Ballings Michel, Dirk Van den Poel, Hespeels Nathalie and Gryp Ruben. (2015). Evaluating multiple classifiers for stock price direction prediction. Expert Systems with Applications. Vol. 42. 7046-7056.

Breiman Leo, Friedman Jan, Stone C. J. and Olshen, R. A. (1984). Classification and regression trees. CRC press.

Donaldson R.Glen and Kamstra Mark. (1996). Forecast combining with neural networks. Journal Forecasting. Vol. 15. 49-61.

Enke David and Thawornwong, Suraphan. (2005). The use of data mining and neural networks for forecasting stock market returns. Expert Systems with Applications. Vol. 29. 927–940.

Hafezia Reza, Shahrabib Jamal and Hadavand Esmaeil. (2011). A bat-neural network multi-agent system (BNNMAS) for stock price pr3ediction: Case study of DAX stock price. Applied Soft Computing. Vol. 29. 196–210.

Han Eui-Hong, Karypis George and Kumar Vipin. (1999). Text categorization using weighted adjusted K-nearest neighbor classification. Technical Report, Department of Computer Science and Engineering, Army HPC Research Centre, University of Minnesota, Minneapolis, USA.

Han Jiawei, Kamber Michelin and Pei Jian. (2006). Data mining: Concepts and techniques, Morgan Kaufmann.

Hassan Md. Rafiul, Nath Barikunth and Kirley Michael. (2007). A fusion model of HMM, ANN and GA for stock market forecasting. Expert Systems with Applications. Vol. 33. Issue 1. 171-180.

Hastie Trevor, Tibshirani Robert and Friedman Jerome. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, New York.

Khemchandani Reshma, Jaya deva and Chandra Suresh. (2009). Knowledge based proximal support vector machines. European Journal of Operational Research.Vol. 195. 914–923.

Kim Eunju, Kim Wooju and Lee Yillbyung. (2003). Combination of multiple classifiers for the customer's purchase behavior prediction. Decision Support Systems. Vol. 34. 167–175.

Lee Ming-chi. (2009). Using support vector machine with a hybrid feature selection method to the stock trend prediction. Expert Systems with Applications.Vol. 36. 10896-10904.

Malkiel Burton G. (2003). The efficient market hypothesis and its critics. The Journal of Economic Perspectives, Vol. 17. No. 1. 59–82.

Osuna Ronta. (2002). Lecture Notes CS: Introduction to pattern recognition. Dayton, Ohio, USA: Wright State University.

Ou, Phichhang and Wang Hengshan. (2009). Prediction of stock market index movement by ten data mining techniques. Modern Applied Science. Vol. 3. No 12. 28-42.

Patel Jigar, Shah Sahil, Thakkar Priyank and Kotecha K. (2015). Predicting stock market index using fusion of machine learning Techniques. Expert Systems with Applications.Vol. 42. 2162–2172.

Quinlan J.R. (1986). Induction of decision trees. Machine Learning. 81-106.

Quinlan J.R. (1993). C4. 5: Programs for machine learning. Morgan Kaufmann. Vol. 1.

Refenes Nicholas, Zapranis Achileas and Francis Gavin. (1994). Stock performance modelling using neural networks: a comparative study with regression models. Neural Networks. Vol. 7.No 2. 375-388.

Tsai Chih-Fong, Lin Yuah-Chiao, Yenb David and Chenc Yan-Min. (2011). Predicting stock returns by classifier ensembles. Applied Soft Computing. Vol. 11. 2452–2459.

Wang Jar-Long and Chan Shu-Hui. (2006). Stock market trading rule discovery using two-layer bias decision tree. Expert Systems with Applications. Vol. 30. 6.5-611.

Xu Xiujuan, Zhou Chunguang and Wang Zhe. (2009). Credit scoring algorithm based on link analysis ranking with support vector machine. Expert Systems with Applications.Vol. 36. 2625–2632.

Yoon Yolsh, Swales Jrmak and Margavio T.M.. (1993). A comparison of discriminate analysis versus artificial neural networks. Journal of Operational Research Society. Vol 8. No 1. 411-420.