ISSN:0975-9646

Clustering-Classification Based Prediction of Stock Market Future Prediction

Abhishek Gupta[#1], Dr. Samidha D. Sharma[*2]

[1,2]*Department of Information Technology,*
NRI Institute of Information Science & Technology Bhopal MP

***Abstract*— Stock market values keeps on changing day by day, so it is very difficult to predict the future value of the market. Although there are various techniques implemented for the prediction of stock market values, but the predicted values are not very accurate and error rate is more. Hence an efficient technique is implemented for the prediction of the stock market values using hybrid combinatorial method of clustering and classification. The dataset is taken from shanghai stock exchange market and is first clustered using K-means clustering algorithm and these clustered values are classified using horizontal partition decision tree.**

***Keywords*—Stock market, K-mean, Prediction.**

I. INTRODUCTION

Stock price forecasting is an important task for investment and financial decision making. Researches and practitioners have given it considerable amount of attention. Stock market is most volatile and highly risky investment. Many approaches have been used for forecasting stock price such as traditional and fundamental methods. Forecasting stock price or financial markets has been one of the biggest challenges to the AI community. A variety of fundamental, statistical, and technical indicators have been offered and used with changeable results But these previous methods have limits and not completely capable to provide accuracy. Data mining and computational intelligence techniques for resolving the problems of stock future price forecasting have become rapidly growing alternative methods for achieving considerable degree of accuracy. The purpose of forecasting research has been largely beyond the capability of traditional AI research which has mainly focused on developing intelligent systems that are supposed to emulate human intelligence [1].

Past approaches to this problem first applied an artificial neural network directly to historical stock data using linear time series modeling [2]. This produced results which over fitted the training data and therefore rendered them unusable in real trading. Additionally to omitting any preprocessing, the neural networks employed only restricted two layers, an output and an input layer. These linear techniques are now known to be provably insufficient for any nonlinear phenomenon including stock price movement.

The rest of this paper is organized as follows: Section 2 is literature survey about various techniques of stock market prediction .Section 3 is proposed methodology that describes k-mean and horizontal partitioning algorithm .Section 4 contains result analysis and finally Section 5 contains conclusion.

II. LITERATURE SURVEY

Chenoweth Tim, Obradovic, and Sauchi [3] relied on a single technical indicator called the average direction index (ADX), which identifies and quantifies trends by averaging the fraction of today's range above or below the earlier day's assortment. The ADX is achieved through a feature selection component and used as input into two separate neural networks (Up and Down) whose results were then polled and applied to a rule base to predict the final market movement. Exclusive of knowing the accurate predictive accurateness, it is complicated to quantitatively judge against the system, which unavoidably comprises rules for trading which may be the actual cause of the monetary gain achieved by the system rather than predictive accuracy.

Roman and Akhtar [4] investigate that if back propagation and recurrent neural networks can be effectively used in designing portfolios across many international stock markets after the trends in these markets for several calendar years are known. Stock prices fluctuate daily resulting in a nonlinear pattern of data.

Shaun and Ruey [5] proposes a stock market forecasting system based on artificial neural network .They train the system with 500 composite indexes of past twenty years .The system produces the forecast and adjust itself by comparing its forecasts with the actual indexes. They also develop a transfer function model to forecast based on the indexes and the forecasts by the artificial neural networks.

Huang, Nakamori, and Shou [6] propose SVM based stock market prediction technique .They compare its performance with Quadratic Discriminant Analysis, Linear Discriminant Analysis and Elman Back-Propagation Neural Network .After comparing them they also propose a combining model by integrating SVM with other classification method.

Mahfoud, Sam, and Mani [7] use genetic algorithms to predict stock prices. Genetic algorithms are encouraged by evolutionary biology and the concept of survival of the fittest. A bulky population of probable algorithmic representations for stock prediction is first produced. Then, each member is implemented and assessed, keeping the algorithms that produce the best results and mixing their properties amongst other high scoring algorithms to obtain a new generation of algorithms in a Darwinian fashion. The process is repetitive until the error has been reduced to an acceptable level.

Naeini, Hamidreza, and Homa [8] propose a neural network based stock market prediction technique .They use two kinds of neural networks a feed forward multilayer perception (MLP) and an Elman recurrent network. They found that MLP neural network is much better in predicting

stock value changes than Elman recurrent network and linear regression method.

Wang, Long, and Chan [9] propose stock prediction technique based on rule discovery. It uses a two-layer bias decision tree. The technique used in this study differs from other studies in two aspects. First, this study modified the decision model into the bias decision model to reduce the classification error. Second, this study uses the two-layer bias decision tree to improve purchasing accuracy. This technique improve purchasing accuracy, investment returns and also have the advantages of fast learning speed, robustness, stability, and generality.

Wu, Yu Lin, and Hsin Lin [10] present a stock trading technique by integrating the filter rule and the decision tree technique. The filter rule is used to generate candidate trading points. These candidate trading points are subsequently clustered and screened by the using a decision tree algorithm C4.5. They apply this technique on Taiwan stock market to justify the technique.

Matsui and Sato [11] proposes a new evaluation method to dissolve the over fitting problem in the GA training. On comparing the conventional and the neighbourhood evaluation they found the new evaluation method performed well than conventional one.

Hadavandi, Shavandi, and Ghanbari[12] proposes an integrated approach for constructing a stock price forecasting expert system using artificial neural networks (ANN) and genetic fuzzy systems (GFS). They use stepwise regression analysis (SRA) to determine factors those have most influence on stock prices. In next stage they divide their raw data into k clusters by means of self-organizing map (SOM) neural networks. At the end all clusters will be fed into independent GFS models with the ability of rule base extraction and data base tuning. By applying this approach on stock price data gathered from IT and Airlines sectors, and compare with previous stock price forecasting methods, they found proposed approach outperforms all previous methods.

Nair, Mohandas, and Sakthivel [13] propose a hybrid decision tree –rough set based system for predicting the next day's trend in BSE-SENSEX.They use technical indicators to extract features from the historical SENSEX data .C4.5 is used to select the relevant features and a rough set based system is applied to induce rules from the extracted features .They compare the performance of hybrid rough set based system with artificial neural network based trend prediction system and the naïve bayes based trend predictor.

Gupta, Aditya, and Dhingra [14] propose a stock market prediction technique based on Hidden Markov Models. In this approach they consider the fractional change in stock value and the intra-day high and low values of the stock to train the continuous HMM.Then this HMM is used to make a Maximum a Posteriori decision over all the possible stock values for the next day. They applied this approach on several stocks, and compare the performance to the existing methods.

Lin, Guo, and Hu.[15] propose a SVM based stock market prediction system .This system selects a good feature subset, evaluates stock indicator and control over fitting on stock market tendency prediction. They tested this approach on Taiwan stock market datasets and found that this system performs well than the conventional stock market prediction system.

Shen, Jiang, and Zhang [16] propose a new prediction algorithm that exploit the temporal among global stock markets and various financial products to predict the next day stock trend with the help of SVM.They applied same algorithm with different regression algorithm to estimate the actual growth in the markets .At last they establish a simple trading model and compare its performance to the existing algorithm.

Kongyu, Wu, and Lin [17] propose a technique for stock price forecasting using genetic algorithm in fuzzy systems to discover rules .The use of genetic algorithm eliminates errors due to noisy data and forms set of rules .After that they apply fuzzy reasoning approach on the rule sets to predict stock market price trends.

III. PROPOSED METHODOLOGY

The proposed methodology implemented here for the prediction of stock markets consists of two stages. First is to apply clustering algorithm such as K-means and then the clustered values are partitioned into number of parties and apply horizontal partition based decision tree algorithm.

A. K-Means Clustering Algorithm

K-means is one of the simplest unsupervised learning algorithms that solve the clustering problem. The algorithm executes in simple way by classifying a given data set through a certain number of clusters (say k clusters) fixed apriori. The basic concept is to define k centers; one for each cluster. These centers should be placed in a tricky way because of different location causes different result. So, it is better way to place them as much as possible far away from each other. In next step it takes each point belonging to a given data set and associates it to the nearest center. When no point remains, the first step is completed and an early grouping is done. Here we need to re-calculate k new centroids as barycenter of the clusters obtained from the previous step. After this there are k new centroids. At this point a new binding has to be done between the same data set points and the nearest new center. A loop is generated. As a result of this loop the k centers change their location step by step until no more changes are done.Finally,this algorithm aims at minimizing an objective function knows as squared error function given by:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} (||x_i - v_j||)^2$$

Where,

'$||x_i - v_j||$' is Euclidean distance between x_i and v_j.'c_i' is the no. of data points in i^{th} cluster.'c' is the no. of cluster centers.

Let $X = \{x_1,x_2,x_3,\ldots\ldots..,x_n\}$ be the set of data points and $V = \{v_1,v_2,\ldots\ldots.,v_c\}$ be the set of centers.

1) Select 'c' cluster centers randomly.

2) Calculate the distance b/w each data point and cluster centers.
3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers..
4) Again calculate the new cluster center using:

$$v_i = \left(\frac{1}{c_i}\right) \sum_{j=1}^{c_i} x_i$$

Where, 'c_i' denotes the number of data points in i^{th} cluster.
5) Again calculate the distance between each data point and new obtained cluster centers.
6) If there is no reassignment of data points then stop, otherwise repeat from step 3.

B. Horizontal Partition Decision Tree

Define n Parties $P_1, P_2 \ldots P_n$. (Horizontally partitioned).
Each Party contains Y set of attributes $A_1, A_2 \ldots A_Y$.
The class attributes C contains c class values $C_1, C_2 \ldots C_c$.
For party P_i where i = 1 to n do
If Y is Empty Then
Return a leaf node with class value
Else if all transaction in T (P_i) have the same class Then
Return a leaf node with the class value
Else
Calculate Expected Information and classify the given sample for each party P_i individually.
Calculate Entropy for each attribute ($A_1, A_2 \ldots A_Y$) of each party P_i.
Calculate Information Gain for each attribute ($A_1, A_2 \ldots A_Y$) of each party P_i.
Calculate Total Information Gain of each attribute of all $P_1, P_2 \ldots P_n$ parties (TotalInformationGain ( )).
$A_{BestAttribute} \leftarrow$ MaxInformationGain ( )
Let $V_1, V_2 \ldots V_m$ be the value of attributes. $A_{BestAttribute}$ partitioned all $P_1, P_2 \ldots P_n$ parties into m parties
$P_1(V_1), P_1(V_2) \ldots P_1(V_m)$
$P_2(V_1), P_2(V_2) \ldots P_2(V_m)$
. .
. .
$P_n(V_1), P_n(V_2), \ldots, P_n(V_m)$
Return the Tree whose Root is labelled $A_{BestAttribute}$ and has m edges labelled $V_1, V_2 \ldots V_m$. Such that for every i the edge Vi goes to the Tree
NPPID3(Y – $A_{BestAttribute}$, C, ($P_1(V_i)$, $P_2(V_i)$, …., $P_n(V_i)$))
End.

IV. RESULT ANALYSIS

The table shown below is the result analysis of the existing and the proposed work. The monthly dataset is taken from Shanghai Stock Exchange and the two techniques when applied on the dataset will predict the following values. The proposed technique provides more close values.

TABLE 1. RESULT ANALYSIS OF THE TWO TECHNIQUES ON MONTHLY DATASET

Monthly dataset	Actual Value	Predicted value by Kongyu Yang et.al.work	Predicted value by Proposed work
1	10.64	9.60	9.94
2	11.66	10.24	10.94
3	12.49	11.46	11.8
4	14.79	13.76	14.1
5	15.22	14.2	14.7
6	13.79	12.75	13.09
7	14.87	13.83	14.17
8	17.05	16.01	16.35
9	19.62	18.6	18.92
10	23.74	22.7	23

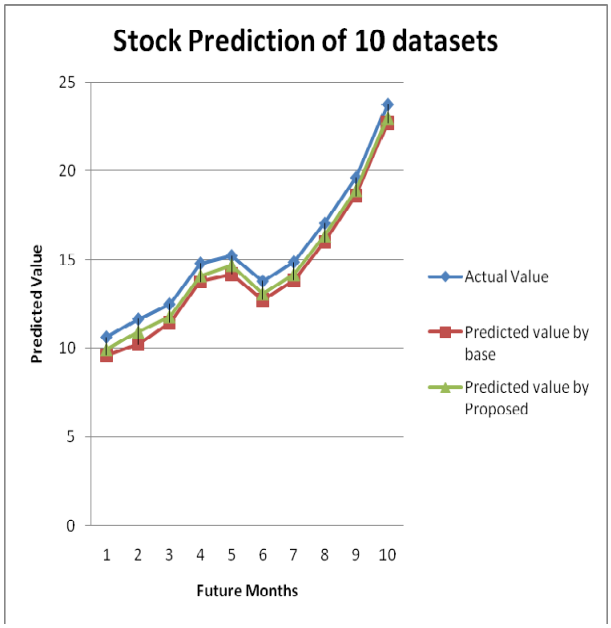

Figure 1. Comparision of the actual and predicted value on monthly dataset

The table shown below is the result analysis of the existing and the proposed work. The yearly dataset is also taken from Shanghai Stock Exchange and the two techniques when applied on the dataset will predict the following values. The proposed technique provides more close values.

TABLE 2. RESULT ANALYSIS OF THE TWO TECHNIQUES ON YEARLY DATASET

Yearly dataset	Actual Value	Predicted value by Kongyu Yang et.al.work	Predicted value by Proposed work
1	2.19	1.97	2.15
2	2.2	1.86	2.11
3	4.01	3.45	3.85
4	4.59	3.75	4.21
5	4.96	4.02	4.68

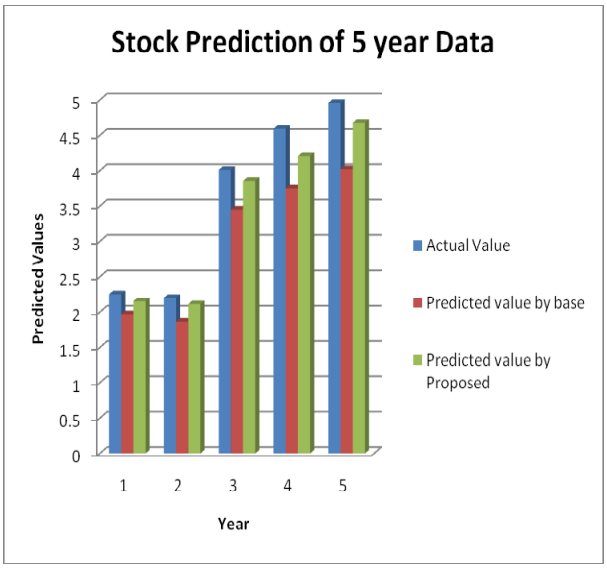

Figure 2. Comparision of the actual and predicted value on yearly dataset

V. CONCLUSION

The proposed technique implemented here for the prediction of stock market provides efficient results as compared to the other existing technique. The proposed methodology provides close prediction of actual value; hence the results will be more accurate and efficient. The algorithms are tested on two datasets, one is monthly and other is yearly dataset. The result analysis shows the performance of the proposed technique.

REFERENCES

[1] Mahdi Pakdaman Naeini, Hamidreza Taremian, Homa Baradaran Hashemi "Stock Market Value Prediction Using Neural Networks", International Conference on Computer Information Systems and Industrial Management Applications (CISIM), pp. 132-136, 2010.

[2] Fischer *Black* and Myron *Scholes* "The Pricing of Options and Corporate Liabilities" The Journal of Political Economy, Vol. 81, No. 3, pp. 637-654, May - Jun *1973.*

[3] Chenoweth, Tim, Zoran Obradovic, and Sauchi Stephen Lee. "Embedding technical analysis into neural network based trading systems." *Applied Artificial Intelligence* 10.6: 523-542, 1996.

[4] Roman, Jovina, and Akhtar Jameel. "Back propagation and recurrent neural networks in financial analysis of multiple stock market returns." *System Sciences, 1996, Proceedings of the Twenty-Ninth Hawaii International Conference on,*. Vol. 2. IEEE, 1996.

[5] Wu, S. and Lu, R. P. "Combining artificial neural networks and statistics for stock market forecasting", Proceedings of the 1993 ACM conference on Computer, pp. 257 – 264, 1993.

[6] Huang, Wei, Yoshiteru Nakamori, and Shou-Yang Wang. "Forecasting stock market movement direction with support vector machine." *Computers & Operations Research* 32, no. 10: 2513-2522, 2005.

[7] Mahfoud, Sam, and Ganesh Mani. "Financial forecasting using genetic algorithms." *Applied Artificial Intelligence* 10, no. 6, 543-566, 1996.

[8] Naeini, Mahdi Pakdaman, Hamidreza Taremian, and Homa Baradaran Hashemi. "Stock market value prediction using neural networks." *Computer Information Systems and Industrial Management Applications (CISIM), 2010 International Conference on*. IEEE, 2010.

[9] Wang, Jar-Long, and Shu-Hui Chan. "Stock market trading rule discovery using two-layer bias decision tree." *Expert Systems with Applications* 30.4, 605-611, 2006.

[10] Wu, Muh-Cherng, Sheng-Yu Lin, and Chia-Hsin Lin. "An effective application of decision tree to stock trading." *Expert Systems with Applications* 31.2: 270-274, 2006.

[11] Matsui, Kazuhiro, and Haruo Sato. "Neighbourhood evaluation in acquiring stock trading strategy using genetic algorithms." *Soft Computing and Pattern Recognition (SoCPaR), 2010 International Conference of*. IEEE, 2010.

[12] Hadavandi, Esmaeil, Hassan Shavandi, and Arash Ghanbari. "Integration of genetic fuzzy systems and artificial neural networks for stock price forecasting." *Knowledge-Based Systems* 23.8, 800-808, 2010.

[13] Nair, Binoy B., V. P. Mohandas, and N. R. Sakthivel. "A decision tree-rough set hybrid system for stock market trend prediction." *International Journal of Computer Applications* 6.9, 1-6, 2010.

[14] Gupta, Aditya, and Bhuwan Dhingra. "Stock market prediction using hidden markov models." *Engineering and Systems (SCES) Students Conference on*, IEEE, 2012.

[15] Lin, Yuling, Haixiang Guo, and Jinglu Hu,"An SVM-based approach for stock market trend prediction." *Neural Networks (IJCNN), The 2013 International Joint Conference on*. IEEE, 2013.

[16] Shen, Shunrong, Haomiao Jiang, and Tongda Zhang.,"Stock Market Forecasting Using Machine Learning Algorithms."2012.

[17] Yang, Kongyu, Min Wu, and Jihui Lin. "The application of fuzzy neural networks in stock price forecasting based On Genetic Algorithm discovering fuzzy rules." *Natural Computation (ICNC), 2012 Eighth International Conference on*. IEEE, 2012.