PREDICTING TREND IN STOCK MARKET EXCHANGE USING MACHINE LEARNING CLASSIFIERS

[1]Khan, W. [2]Ghazanfar, M. A. [3]Asam, M. [4]Iqbal, A. [5]Ahmad, S. [6*]Javed Ali Khan
[1,2,3,4,5]Software Engineering Department, University of Engineering and Technology, Taxila
[6]Software Engineering Department, University of Science and Technology, Bannu
*Correspondence E-mail: engr_javed501@yahoo.com

ABSTRACT: *Prediction in stock market is an interesting and challenging research topic in machine learning. A large research has been conducted for prediction in stock market by using different machine learning classifiers. This research paper presents a detail study on data of London, New York, and Karachi stock exchange markets to predict the future trend in these stock exchange markets. In this study, we have applied machine learning classifiers before and after applying principle component analysis (PCA) and reported errors and accuracy of the algorithms before and after applying PCA. The performance of the selected algorithms has been compared using accuracy measure over the selected datasets.*

Keywords: SVM, KNN, Machine Learning, Stock Market Prediction, Naïve Bayes classifier.

INTRODUCTION

The stock market is an evolutionary, complex and a dynamic system. Market prediction is characterized by noise, data intensity, non-stationary, uncertainty and hidden relationships [5]. The prediction of trend in stock market exchange has been a challenging and important research topic. It is challenging because the data is noisy and not stationary. It is important because it can yield important results for decision makers. Stock market is such a location where companies invest high capital and do their shares trading. Stock market prediction has disproved the Efficient Market Hypothesis which states that it is impossible to predict the market because it is efficient. Researchers have proved that it is possible to predict the stock market. The ability of making future stock market prediction is an important factor for investors for making money. It also helps investors to make selling or buying decisions to generate higher profits.

In this paper, we have predicted future trend in the stock exchanges of Karachi, New York and London before and after applying PCA and compared the accuracy of the algorithms before and after applying PCA by using different matrices like root mean square error (RMSE), mean absolute error (MAE) and accuracy.

The rest of the paper is organized as follows. In part II, we discuss some prior related work. Part III illustrates some machine learning classifiers used in this research. Part IV illustrates our experimental method and the datasets used. We present our experimental results in part V and in part VI we conclude and suggest some future work.

RELATED RESEARCH

Researchers have employed different machine learning classifiers for the prediction of stock market. This prediction is based on previous records called training data set. Normally 80% of the data set comprises the training data set. The data set that is tested on the trained classifier is called the testing dataset that comprises 20% of the data. We have used data from 2009 to 2013 as the training data and data of 2014 as the testing data in our selected datasets. Recently, prediction has been recognized as an important topic in Machine Learning. Eclipse was used as the tool to realize our research results in java programming language.

A number of machine learning and artificial intelligence techniques have been used for the prediction of stock market. Kumkum Garg in [1] proposed a hybrid machine learning system that was based on Support Vector Machine (SVM) and Genetic Algorithm (GA) for the prediction of stock market. Mittermayer in [2] implemented News CATS system to predict trends in stock price. This system used press releases data for price prediction for the time after press releases.

Robert P, Schumaker and H Chen [3] used different textual representations to find financial related news articles and quotes. This approach was then applied to predict discrete stock price after the release of the news article. Kyoung-jae and Kim [4] applied SVM to predict stock market index and examined the feasibility of applying SVM in financial prediction by comparing SVM with artificial neural networks (ANNs) and case based reasoning (CBR).

Wang *et al.* [5] predicted the direction of financial movement with SVM. SVM was compared with other techniques and it was found that SVM outperformed the other classifiers. Wohar *et al.* [vi] performed a detailed analysis of in-sample and out-of-sample tests predictability of stock return in order to better understand the evidence on return predictability.

Wuthrich *et al.* [6] predicted stock market using information that is contained in articles published on the Internet. They used textual information because in addition to numeric data, it increases the quality of input data for prediction. Yu-Kun *et al.* [8] used neural network technique called FSVMR (fuzzy support vector machines regression) for stock index forecasting. The objective was to improve FSVMR accuracy in terms of pre-processing of data, parameters selection and kernel function selection.

Fenghua *et al.* [9] used singular spectrum analysis (SSA) to decompose price series of stock into the terms of market fluctuation, trend and noise to get more accurate results. These terms were then introduced into the SVM for price prediction. Kara *et al.* [10] developed two models that were based on ANN and SVM and compared their performances in predicting direction of movement in Istanbul stock exchange index.

L. Cao and Francis E.H. Tay *et al.* [11] examined the feasibility of applying SVM in financial prediction and investigated the functional properties of SVM in this prediction. These functional properties were obtained using selecting free parameters of SVM. Kim and Han *et al.* [12] proposed a hybrid model of ANN and GA for discretization of features to predict stock price index. In this research, GA

was used to improve the learning algorithm and reduce the complexity in feature space. Features discretization was conducted to simplify the learning process and improve generalizability of the learned results.

Tsai *et al.* [13] used a hybrid system called SOFM-SVR that used support vector regression (SVR) and self organizing feature map (SOFM) technique and a filter based feature selection to reduce training time cost and to improve prediction accuracies. Zeng *et al.* [14] investigated that public sentiment expressed as collections of daily Twitter posts can be used for the prediction of stock market.

Machine Learning Classifiers employed for forecasting stock exchange data

Dase and Pawar [15] performed a literature review and found that ANN is very useful to predict the world stock market. L. J. Cao and Francis [16] applied SVM for forecasting in financial time series data. The feasibility of application of SVM to financial time series forecasting was examined by comparing SVM with multi layer BP (back propagation) NN and the regularized RBF (Radial Basis Function) NN. It was found that SVM outperformed BP NN and there was comparable generalization performance between SVM and RBF NN in financial forecasting.

Zhang *et al.* [17] proposed a new machine learning prediction algorithm, SVM that exploit temporal correlation among stock markets and different financial products to find the next day stock trend. Salam *et al.* [18] proposed a hybrid algorithm that integrated Particle swarm optimization (PSO) and least square SVM (LS-SVM). In this algorithm, PSO was used to select best parameters for LS-SVM to optimize LS-SVM to predict the daily stock prices.

Jimoh *et al.* [19] used regression analysis for stock price prediction. Stock prices were extracted from stock exchange official list. These prices were used to build a database and variable values were extracted from this database. This data was then used to predict financial market price.In this section, we briefly discuss machine learning classifiers used for stock exchange prediction in this research.

Support Vector Machine

To predict the future trends in stock market exchange, we used Support Vector Machine (SVM). SVM is a machine learning pattern classification algorithm proposed by Vapnik and co-workers [20]. It is an important algorithm in machine learning because of its performance in terms of accuracy as compared to other classifiers of machine learning [21]. Unlike old algorithms which reduce the training error, SVM tries to minimize generalization error by increasing the margins between the data and the separating hyper plane. SVM is very attractive algorithm due to its property of reducing information in training data and providing sparseness by using very small number of Support Vectors for finding class label.

SVM is a supervised learning technique that minimizes classification error and maximizes the geometric margins. Therefore it is also called maximum margin classifier. A maximum separating hyper plane is created in high dimensional featured space. There are two parallel hyper planes that separate the data and are created on both sides of the hyper plane. The separating hyper plane maximizes the distance between these two parallel hyper planes. If the margin between these parallel hyper planes is larger, then the generalization error of classifier will be better.

Naïve Bayes Classifier

Naïve Bayes classifier is a statistical classifier as it can predict probabilities of a class membership. Naïve Bayes classifier is based on Bayes theorem. It makes use of class conditional independence. According to this conditional independence, attributes are independent of each other given the class. It works as follows:

Suppose T represents the set of all training examples each with their own class labels i.e. there are k classes c_1,c_2, c_3,…,c_k. Each example is denoted by n-dimensional vector, V={v_1,v_2,v_3,…,v_n}, showing n measured values of the n attributes a_1,a_2,a_3,...,a_n respectively. Given an example D, this classifier will predict that D belongs to the class that has the highest aposteriori probability conditioned on D. Therefore we find the class that maximizes P (ci/D). According to Bayes Theorem, we have:

P(ci/ D) = P (D|ci) P (ci)/ P (D) (1)

Here P(D) is called the normalization constant and can be ignored as it has same value for all given classes. If we ignore apriori probabilities, P (ci), then Equation (1) is called maximum likelihood. The class apriori probabilities can easily be found by (ci) = $freq$ (ci, T) / |T|; however, the estimation of computing (D|ci) is very costly when there are large attributes given. This problem is overcome by Naïve Bayes by making class conditional *independence assumption*. Mathematically it can be written as:

(D|c_i)= $\prod_i^n (d_i|c_i)$. (2)

The probabilities ($d_1|c_i$), ($d_2|c_i$),...,($d_n|c_i$) are computed from the training set. In Equation 2, the term d_k denotes the value of attribute a_k for the given sample. When a_k is categorical, then ($d_k|c_i$) is the number of examples of class c_i in T having the value d_k for attribute a_k, divided by the number of examples of the class c_i in T (i.e. frequency (c_i, T)).

K Nearest Neighbors

K Nearest Neighbor (KNN) is a simpler classifier of machine learning. The problem of stock prediction is mapped into classification which is based on similarity. The stock training and testing data is mapped into vectors. These vectors represents dimensions of features. Eucledian distance is measured for making decisions. Following are steps of KNN for making prediction in stock market:

1. Find K i.e nearest neighbors. We have used value of 12 for K.
2. Distance is measured between training and testing examples.
3. Training data is sorted based on measured distances.
4. Majority vote is used for the K nearest neighbor class labels.
5. Use the majority vote as prediction value for the Query record.

METHODOLOGY

In this section, we are discussing our proposed research methodology. Figure 01 shows this research methodology.

Datasets Selection

For the evaluation of the algorithm and finding future trends in our proposed research, we have used three datasets of stock exchanges of Karachi, London and New York and performed analysis of all the datasets. The datasets were separated into training and testing datasets. These datasets contain historical data from 2009 to 2014. We have used data from 2009 to 2013 as training data and data of 2014 as testing data. As our target is to predict the trend in upcoming years, therefore the selected data contain historical information of stock exchange of 5 years. We have conducted 2-fold cross validation over the training dataset to learn the optimal parameters.

Features Selection

Our selected datasets has a total of 9 common input features including *Date, Open, High, Low, Close, Volume, Trend, Sentiment and Future Trend Value.*

Features Reduction by applying PCA

After applying the classifier on all the features, we reduced the input features of the dataset by applying principal component analysis (PCA) and then applied the classifiers over the reduced features. We used five out of nine features of the datasets. We found the results after features reduction as well.

Finding Accuracy

There are different measures that are used to measure prediction accuracy like Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and accuracy. Their mathematical formulation is given as:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(a_i - p_i)^2} \quad (3)$$

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|a_i - p_i| \quad (4)$$

$$\text{Accuracy} = {}^{N_c}/_{N} \quad (5)$$

Here, *N* represents the total number of records predicted, *ai* represents the actual value of a record and *pi* represents the predicting value of a record and *Nc* denotes the number of records that are correctly classified.

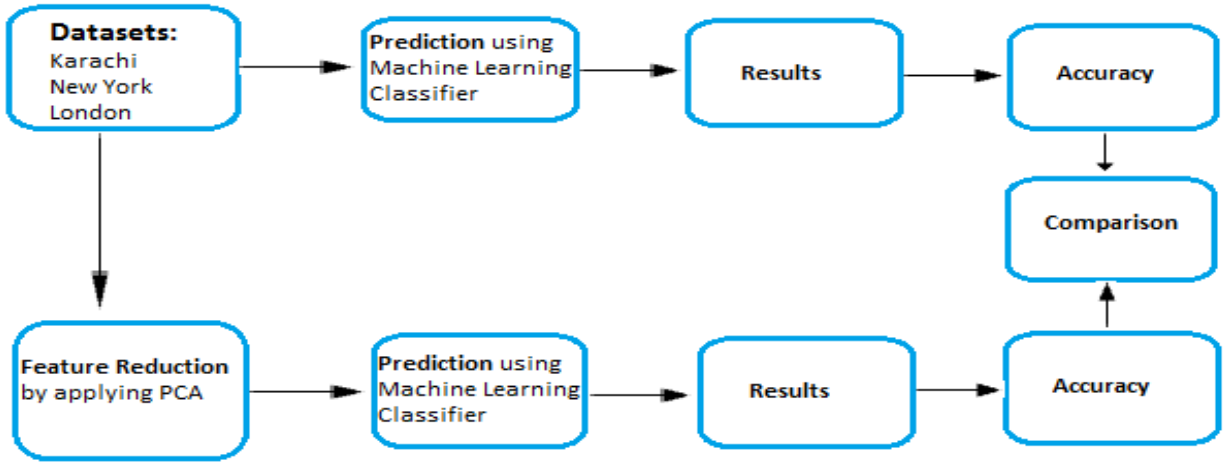

Fig. 1. Block diagram of the proposed research approach

RESULTS AND DISCUSSION

The experimental results are shown in Fig. 2(a-b) and Fig. 3(a-b). We have repeated the experiment on the three selected datasets (*KSE: Karachi Stock Exchange, LSE: London Stock Exchange, NSE: New York Stock Exchange*) before and after applying PCA for all the classifiers selected and have calculated RMSE, MAE and Accuracy against each classifier as shown in Table I and Table II. The horizontal axis shows the datasets while the vertical axis shows Accuracy in Fig. 2 and MAE in Fig. 3. The lowest MAE shows the performance of that algorithm for prediction. It is observed from our experimental results and graphs that KNN has highest Accuracy and lowest MAE and it shows that it is the best classifier in terms of accuracy and MAE for prediction as compared to SVM and Naïve Bayes. Our experimental results also show that SVM is a second best classifier for stock market prediction after KNN in terms of high accuracy and low MAE as compared to Naïve Bayes.

We have applied PCA over the selected datasets and repeated the experiment over reduced data. Our results show that accuracy increases while MAE decreases when the classifiers are applied over the reduced datasets. This effect is shown in Fig. 2(b) and Fig. 3(b).

Prediction in stock market is a potential and a beneficial area of research for business decision makers. This paper attempts to predict stock market prediction in Karachi, New York and London stock exchanges. In this study, we analyzed historical data of these datasets taken over a period of five years.

Different classifiers of machine learning have been used to predict future trend in these datasets. KNN have shown good performance in terms of Accuracy and MAE.

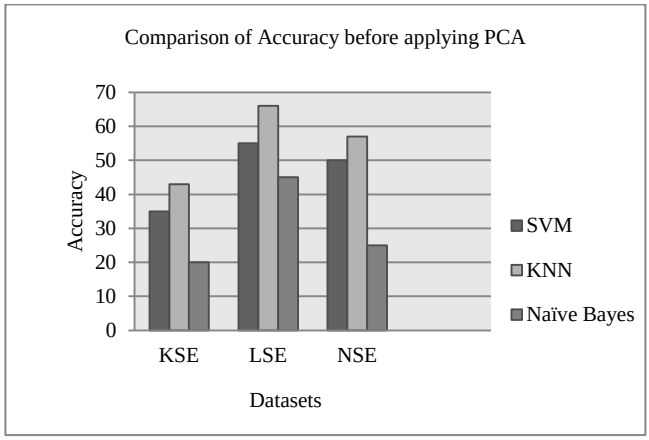

Fig. 2(a). Comparison of Accuracy before applying PCA.

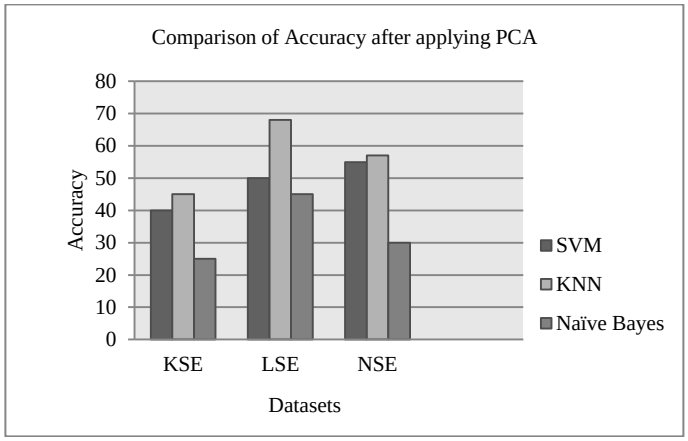

Fig. 2(b). Comparison of Accuracy after applying PCA

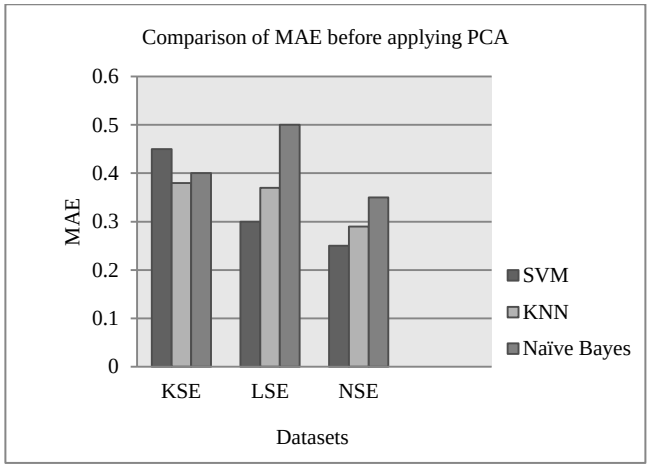

Fig. 3(a). Comparison of MAE before applying PCA

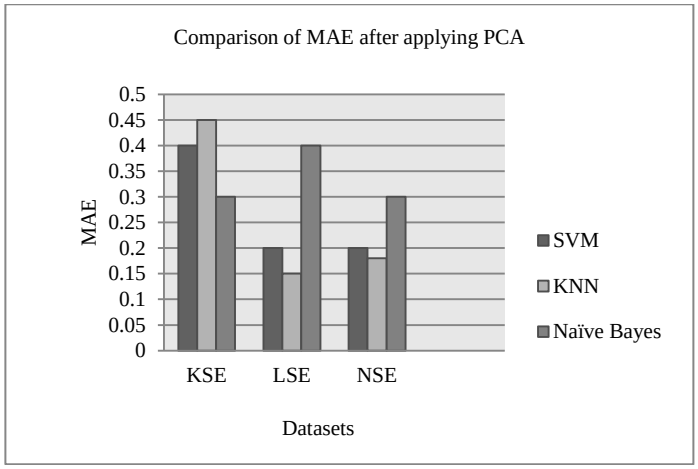

Fig. 3(b). Comparison of MAE after applying PCA

Table 1: and Table 2 show details of our experimental results.

Table 1: Summary of RMSE, MAE and accuracy before applying PCA for three classifiers

Datasets	MAE	RMSE	Accuracy (%)
KNN			
KSE	0.38	0.40	43
LSE	0.37	0.31	66
NSE	0.29	0.3	57
SVM			
KSE	0.45	0.47	35
LSE	0.3	0.33	55
NSE	0.25	0.28	50
Naïve Bayes			
KSE	0.4	0.45	20
LSE	0.5	0.55	45
NSE	0.35	0.4	25

Table 2: Summary of RMSE, MAE and accuracy after applying PCA for three classifiers

Datasets	MAE	RMSE	Accuracy (%)
KNN			
KSE	0.45	0.37	45
LSE	0.15	0.25	68
NSE	0.18	0.23	57
SVM			
KSE	0.4	0.42	40
LSE	0.2	0.3	50
NSE	0.2	0.24	55
Naïve Bayes			
KSE	0.3	0.4	25
LSE	0.4	.45	45
NSE	0.3	0.36	30

CONCLUSION AND FUTURE WORK

As our future work, we consider social media data in addition to historical data in stock market prediction. The social media data will improve our prediction results.

REFERENCES

1. Choudhry, Rohit, and Kumkum Garg. "A hybrid machine learning system for stock market forecasting."

World Academy of Science, Engineering and
2. Mittermayer, Marc-André. "Forecasting intraday stock price trends with text mining techniques." System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference on. IEEE, 2004.
3. Schumaker, Robert P., H. Chen. "Textual analysis of stock market prediction using breaking financial news: The AZFin text system." ACM Transactions on Information Systems (TOIS) Volume 27, Issue 2, pp: 12, 2009.
4. Kim, Kyoung-jae. "Financial time series forecasting using support vector machines." Neurocomputing Volume 55, Issue 1, pp:307-319, 2003
5. Huang, Wei, Y. Nakamori, S. Wang. "Forecasting stock market movement direction with support vector machine." Computers & Operations Research, Volume 32, Issue 10, pp: 2513-2522, 2005.
6. D. E. Rapach, M. E. Wohar. "In-sample vs. out-of-sample tests of stock return predictability in the context of data mining." Journal of Empirical Finance, Volume 13, Issue 2, pp: 231-247, 2006.
7. B. Wuthrich,V. Cho, S, Leung, J. Zhang, W. Lam. "Daily stock market forecast from textual web data." Systems, Man, and Cybernetics. 1998 IEEE International Conference. Volume 3, 1998.
8. Bao, Yu-Kun, Liu, Guo, Wang "Forecasting stock composite index by fuzzy support vector machines regression." Machine Learning and Cybernetics. Proceedings of 2005 International Conference on. Vol. 6. IEEE, 2005.
9. Fenghua, W. E. N., Jihong, X.I.A.O, Zhifang. "Stock Price Prediction based on SSA and SVM." *Procedia Computer Science* 31 (2014): 625-631.
10. Kara, Yakup, M. A. Boyacioglu, Ö. K. Baykan. "Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange." Expert systems with Applications, Volume 38, Issue 5,pp: 5311-5319,2011.
11. Cao, Lijuan, and Francis EH Tay. "Financial forecasting using support vector machines." Neural Computing & Applications, Volume 10, Issue , pp: 184-192,2001

Technology Volume 39 Issue. 3 pp: 315-31, 2008.
12. Kim, Kyoung-jae, and I. Han. "Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index." Expert systems with Applications, Volume 19, Issue 2,pp: 125-132,2000.
13. Huang, Cheng-Lung, C. Tsai. "A hybrid SOFM-SVR with a filter-based feature selection for stock market forecasting." Expert Systems with Applications, Volume 36, Issue 2, pp: 1529-1539,2009.
14. Bollen, Johan, H. Mao, X. Zeng. "Twitter mood predicts the stock market." Journal of Computational Science, Volume 2, Issue 1 , pp: 1-8,2011.
15. Dase R. K., Pawar D. D. "Application of Artificial Neural Network for stock market predictions: A review of literature." International Journal of Machine Intelligence, Volume 2, Issue 2, pp: 14-17, 2010.
16. L.J. Cao, F. E.H. Tay. "Support vector machine with adaptive parameters in financial time series forecasting." Neural Networks, IEEE Transactions on 14.6 pp: 1506-1518,2003.
17. Shen, Shunrong, H. Jiang, T. Zhang. "Stock market forecasting using machine learning algorithms." , 2012.
18. Hegazy, Osman, O. S. Soliman, M. A. Salam. "A Machine Learning Model for Stock Market Prediction." arXiv preprint arXiv, pp:1402.7351, 2014.
19. Olaniyi, S. A. Sulaiman, K. S. Adewole, R. G. Jimoh. "Stock trend prediction using regression analysis–a data mining approach." ARPN Journal of Systems and Software, Volume 1, Issue 4 ,pp: 154-157, 2011.
20. Boser, B., Guyon, I., Vapnik, V. "A training algorithm for optimal Margin classifiers." Fifth Annual Workshop on Computational Learning Theory, New York: ACM Press 1992.
21. Srivastava, Durgesh K., Lekha Bhambhu. "Data classification using support vector machine." ,2010.