2017 13th International Conference on Computational Intelligence and Security

Prediction of Stock Market by Principal Component Analysis

Muhammad Waqar, Hassan Dawood,Muhammad Bilal Shahnawaz, Mustansar Ali Ghazanfar
Software Engineering Department
University of Engineering and Technology
Taxila, Pakistan
mwaqar588@gmail.com,
Hassan.dawood@uettaxila.edu.pk,bilalbajwa3840@gmail.com, mustansar.ali@uettaxila.edu.pk

Ping Guo
Image Processing and Pattern Recognition Laboratory
Beijing Normal University
Beijing, China
pguo@bnu.edu.cn

***Abstract*—The categorization of high dimensional data present a fascinating challenge to machine learning models as frequent number of highly correlated dimensions or attributes can affect the accuracy of classification model. In this paper, the problem of high dimensionality of stock exchange is investigated to predict the market trends by applying the principal component analysis (PCA) with linear regression. PCA can help to improve the predictive performance of machine learning methods while reducing the redundancy among the data. Experiments are carried out on a high dimensional spectral of 3 stock exchanges such as: New York Stock Exchange, London Stock Exchange and Karachi Stock Exchange. The accuracy of linear regression classification model is compared before and after applying PCA. The experiments show that PCA can improve the performance of machine learning in general if and only if relative correlation among input features is investigated and careful selection is done while choosing principal components. Root mean square error (RMSE) is used as an evaluation metric to evaluate the classification model.**

Keywords- principal component analysis, stock exchange prediction, linear regression, root mean sqaure error

I. INTRODUCTION

Prediction is a process to make assumption of future based on existing data. The more precise the prediction, the more it could be easier to make decision for future. Prediction of stock exchange trends has been an interesting topic in the field of pattern recognition and machine learning because of its possible monetary profit. A stock market is an organized set-up with a regulatory body and has registered members who can buy or sell shares. It's a public market, where different companies invest high capital and do trading of their shares. Stock market prediction provides information about stock market, which can help the shareholders to make decision about trading. It may serve as warning system for long-term shareholders while for short-term investors may serve as recommender system.

One of the problems with Stock market data is that the available data is highly volatile with very high dimensions. Many of the attributes are highly correlated and makes prediction of stock market a highly challenging, complicated and daunting task. The best solution for above problem is to reduce dimensions of data. Measuring correlation between data dimensions can do reduction and discarding those attributes, or dimensions which have least impact on overall prediction model. Principal component analysis is one of effective technique, which can be applied to reduce the dimensionality of data.

In this paper, we have predicted the trend of three stock exchanges by using linear regression as a classification model. The past values are used to build a classification function and then uses this function to predict about future values. Principal component analysis is used with linear regression model to check that whether PCA has improved the accuracy of model or not.

This paper has been organized in V sections: Section I is introduction. Section II briefly describes the literature review. Section III contains proposed methodology while section IV presents experimental results. Conclusion of the study has been discussed in section V.

II. LITRATURE REVIEW

David Enke and Suraphan Thawornwong [1] employed machine learning techniques to assess the predictive relationship of versatile financial and economic variables by using neural network models for effective forecast of future values. Mayankkumar B Patel *et al*. [2] used artificial neural network (ANN) to make prediction of stock price for the companies listed under National Stock Exchange (NSE). Dase R.K and Pawar D.D. [3] used neural network to predict the stock rate as it possesses the ability to extract utile information from a large dataset.

Halbert white [4] used neural network and learning methods to model nonlinear regularities in asset price movement and reported his findings. Debashish Das *et al*. [5] applied a combination of neural network and data mining to make a reliable prediction of stock market because neural network is capable to extract utile information from a large dataset and data mining has ability to forecast future trends. Sneha Soni [6] did survey on different machine learning techniques that were used to make stock market prediction and identified Artificial Neural Network as a dominant technique to forecast stock market.

0-7695-6341-4/17/31.00 ©2017 IEEE
DOI 10.1109/CIS.2017.00139

Shunrong Shen and Tongda Zhang [7] proposed an algorithm which employed the notion of temporal correlation among numerous products and global markets to make prediction of next day stock market trend. Prediction accuracies were measured using Support Vector Machine (SVM) algorithm. Yanshan Wang *et al.* [8] proposed an empirical study on the Korean and Hong Kong stock market. Principal Component Analysis was used for features selection and Support Vector Machine for stock market prediction. Wen Fenghua *et al.* [9] used singular spectrum analysis (SSA) to decompose stock price in diverse terms with varying economic features and made price prediction by introducing these features into the support vector machine (SVM).

Rohit Choudhry *et al.* [10] proposed a hybrid machine learning system to predict stock prices by using correlation among stock prices of different companies. Hybrid system was based on Genetic Algorithm (GA) and Support Vector Machine (SVM). Karazmodeh *et al.* [11] employed Particle Swarm Optimization [PSO] and Support Vector Machine to propose a computationally more efficient model to predict stock price by using correlation among stock prices of various companies.

Abdulsalam sulaiman *et al.* [12] built a database of price lists officially published by Nigerian Stock Exchange and applied data mining to uncover hidden patterns, relationships among various variables and to extract values of variables. Moving average method was employed on these extracted values to make prediction of future stock market prices. Phichhang Ou *et al.* [13] applied ten different data mining techniques to predict price movement of Hang Seng index of Hong Kong stock market.

In existing literature there is a limited work that has been done to investigate the effect of PCA on relative accuracy of classification models. In general, PCA aims to reduce dimensions of dataset so that redundancy can be removed. The drawback to this approach is that sometimes while reducing dimensions valuable date losses, which can effect overall quality of predictions. We have investigated the effect of PCA application on three stock market datasets and analyzed the relative accuracy of our classification model. Some effort also has been done to investigate the input features effect on performance of PCA. For this, some input features are extracted from dataset.

III. Classification Approach Employed for preciting Stock Exchange data

A. Linear Regression

Linear regression is a linear approach used to express the relationship between two variables; one is called independent variable X which can be more than one and the other scalar variable Y dependent on X. In linear regression, relationships between variables are modeled using linear predictor functions and values for unknown variables are quantified using the known variables. Such models are called linear models. [14]. Regression starts with inclusion of a large data set and a predictive model can be built by training this data. In the training process, a regression algorithm estimates the value of concerned variable as a function of the predictors for each case and model is developed which can predict the output value of the target variable for some new input data. This can also be employed on a different data set to estimate the unknown target values. We are given input in the form of (X_i, Y_i), where i =1 to n, and we have to predict Y_{n+1} for a new point X_{n+1}. The equation for linear regression can be expressed as:

$$Y_{n+1} = wt{*}x_{n+1} \tag{1}$$

Where w is a parameter that is to be estimated.

B. Principal Component Analysis

Principal Component Analysis (PCA) is a method used to minimize dimensions of different variables in a given data set. It generally involves application of covariance analysis among different attributes. The original data is mapped into a new coordinate system based on the variance within given data. PCA employs a mathematical procedure to transmute correlated variables into linearly uncorrelated variables called principal components. To reduce dimensionality of transformed data, only first few components are considered because the first principal component account for the largest variance in data and each succeeding component accounts for as much of the remaining variability as possible.

PCA is useful when there are a large number of correlated dimensions that often contain a lot of data redundancy. PCA can be employed to reduce this redundancy, which results in reduction of highly correlated data into small number of un-correlated principal components that actually account for most of the variance in the highly correlated data.

IV. Experimental setup

Experiments have been conducted on three different stock exchanges data sets: London stock exchange (LSE), New York stock exchange (NYSE) and Karachi stock exchange (KSE). The selected data has total 5 input features including open, High, Low, Current and Change. The main focus of the research is to predict the stock market trend (either will increase or decrease). Therefore, the change of a feature over time is more important than the absolute value of each feature. A new feature is defined as “differ” which is actually the difference between stock market’s closing date and opening date values. Along with trend, mid-month trend, monthly trend is also added to the input features in order to obtain maximum correlation between input data dimensions so that the effect of PCA analysis on highly correlated input data can be examined, effectively. The main aim is to obtain maximum correlation between features so that principal components extraction for PCA can become easier. To judge the performance and accuracy of linear regression on provided datasets, root mean square error is used as evaluation criteria which is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(d_i - z_i)^2} \tag{2}$$

Where N denotes the total number of samples forecasted, di denotes the actual value of a sample, z_i denotes the forecasting value of a sample. The value of both z_i and d_i can be ternary i.e. these attributes can have any of three values: -1, 0 and 1. The possible value of input attributes like trend, differ, monthly trend etc. can have positive or negative value as well.

TABLE I. COMPARISON BETWEEN ROOT MEAN SQUARE ERRORS.

Stock Exchange	Root Mean Square Error
London Stock Exchange	16.43
New York Stock Exchange	36
Karachi Stock Exchange	0.13

From table I, it can be concluded that KSE has minimum root mean square error and has highest accuracy while NYSE has highest root mean square error with lowest accuracy. After this PCA is applied on input data to reduce dimensionality of datasets. PCA does not use the original input space vectors to rank its features, however attempts to combine most similar ones of those input vectors and form a new feature space vectors. As a result of this new combination, the new class decisions are based on reduced set of input vector space. The ranking of features is accomplished according to new modified class decisions and highly ranked features are now used to build classification model.

The comparison results of RMSE before and after applying PCA are shown in Table 2. It is clear that in case of LSE and NYSE, RMSE decreases significantly and the accuracy is increased after applying PCA. While in case of KSE, RMSE increases from 0.1 to 1.01 showing that there is relative decrease in accuracy after applying PCA. The increase in accuracy for LSE and NYSE can be explained by the fact that PCA reduces the dimensions by keeping only as many eigenvectors as needed to explain 99% of the variance. So, by reducing dimensions and irrelevant attributes accuracy of models is increased while decrement in accuracy of KSE can be explained by the fact that sometimes application of PCA results in loss of some critical information that could be important for classification and causes decrement in classification accuracy. Also, it often occurs that increasing number of principal components while applying PCA introduces some error in classification that can decrease the accuracy of classification. From results, it can be observed that appropriate selection of input features can greatly enhance the overall performance of PCA. So, selection of appropriate principal components are vital for enhancing PCA's utility.

TABLE II. COMPARISON BETWEEN ROOT MEAN SQUARE ERRORS

Stock Exchange	RMSE Without PCA	RMSE With PCA
London Stock Exchange	16.43	1.4
New York Stock Exchange	36	1.00
Karachi Stock Exchange	0.13	1.01

Figure.1 shows the starting RMSE values of LSE, NYSE and KSE before the application of PCA and respective RMSE values after the application of principal component analysis.

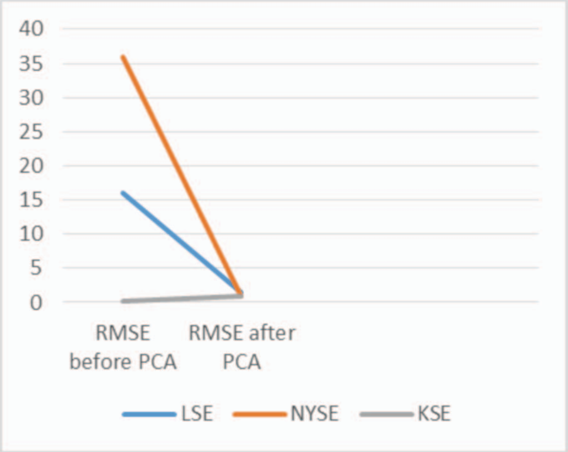

Figure 1. Graphical comparison of RMSE before and after applying PCA.

$$\text{LRM Trend} = (-0.0911) * \text{Differ} + (-1.1834) \qquad (3)$$

Equation (3) shows the linear regression function that is used for stock market prediction.

V. CONCLUSION

This paper has underlined the utilization of PCA to improve the performance of machine learning model in classification of high dimensional data. We have investigated the effect of input features on overall quality of PCA. It is verified that PCA always does not guarantee for improved accuracy. Sometime through the use of PCA, some low errors can be achieved despite a major reduction of input data which may result an overall reduction in accuracy of classification model. That can be viewed in case of PCA implementation on KSE dataset in presented research work. This paper can be used as a base line in future that researchers can use to understand the effect of implementing PCA on highly correlated datasets. It can also help researchers for selecting optimal principal components when applying PCA on a particular classification model having high dimensional input dataset and can also be used when investigating the automatic selection of parameters for techniques, such as the number of PCs, kernel parameters for SVM and k for k-NN.

VI. ACKNOWLEDGMENT

This work is fully supported by the grants from the National Science Foundation of China (61375045), Beijing Natural Science Foundation (4142030) and the Joint Research Fund in Astronomy (U1531242) under cooperative agreement between the National Natural Science Foundation of China (NSFC) and Chinese Academy of Sciences (CAS).

REFERENCES

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

[2] M. B. Patel and S. R. Yalamalle, "Stock Price Prediction Using Artificial Neural Network" International Journal of Innovative Research in Science, Engineering and Technology, vol. 3, pp.13755-13762, June 2014.

[3] R.K. Dase and D.D. Pawar, "Application of Artificial Neural Network for stock market predictions: A review of literature," International Journal of Machine Intelligence, vol. 2, pp. 14-17, 2010.

[4] H. White, "Economic prediction using neural networks: the case of IBM daily stock returns" IEEE International Conference on Neural Networks (IEEE 98), San Diago, IEEE Press, July 1998, pp.451-459, doi:10.1109/ICNN.1988.23959.

[5] D. Das and M.S. Uddin, "Data Mining and Neural Network techniques in Stock market prediction: A methodological review," International Journal of Artificial Intelligence & Applications (IJAIA), vol.4, pp. 117-127, January 2013.

[6] S. Soni, "Applications of ANNs in Stock Market Prediction: A Survey," International Journal of Computer Science & Engineering Technology (IJCSET), vol. 2, pp.71-83. 2013.

[7] Shunrong Shen, Haomiao Jiang, and Tongda Zhang. "Stock Market Forecasting Using Machine Learning Algorithms," International Journal of Machine Intelligence, vol. 3, pp. 17-22, 2012.

[8] Y. Wang and I. C. Choi, "Market Index and Stock Price Direction Prediction using Machine Learning Techniques: An empirical study on the KOSPI and HIS," International Journal of Business Intelligence and Data Mining, vol.1, pp.1-13, September 2013.

[9] W. Fenghuaa, X. Jihong, H. Zhifang and G. Xu. "Stock Price Prediction based on SSA and SVM." Proc. 2nd International Conference on Information Technology and Quantitative Management (ITQM 14), Elsevier, 2014, pp. 645-631, doi: 10.1016/j.procs.2014.05.309.

[10] R. Choudhry and K. Garg, "A Hybrid Machine Learning System for Stock Market Forecasting," International Scholarly and Scientific Research & Innovation, vol.2, pp.242-245, 2008.

[11] M. Karazmodeh, S. Nasiri and S. Majid Hashemi, "Stock Price Forecasting using Support Vector Machines and Improved Particle Swarm Optimization," Journal of Automation and Control Engineering, vol. 1, pp.173-176, June 2013.

[12] A. S. Olaniyi, A. kayode, R.G. Jimoh, "Stock Trend Prediction using Regression Analysis – A Data Mining Approach," ARPN Journal of Systems and Software, vol.1, pp. 154-157, July 2010.

[13] P. Ou and H. Wang, "Prediction of Stock Market Index Movement by Ten Data Mining Techniques," Modern Applied Science, vol. 3, pp.28-42, December 2009.

[14] "Linear Regression," available on-line at https://en.wikipedia.org/wiki/Linear_regression, 2015