Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 05, May - 2017

Performance Analysis of Data Mining Techniques for Stock Price Forecasting

Sarika Shrivastava
Dept. of IT, Dr. C.V. Raman University,
Bilaspur (C.G.), India

A. K. Shrivas
Dept. of IT, Dr. C. V. Raman University,
Bilaspur (C.G.), India

***Abstract*: Prediction of stock price is very challenging task due to dynamic changes value of company stock. The successfully stock prediction helps to profit the company otherwise face the problem of loss. This paper presents the analytical study of various data mining based predictive models to predict the stock in financial domain and find the most consistent prediction model among them. In Stock Market, Data mining play very important role to analysis of data. Data mining techniques can be applied on past and present financial data to generate pattern and decision making. In this paper, we have used predictive techniques like Decision Tree (DT), Random Forest(RF), Support Vector Machine (SVM), Random Tree(RT) and Multilayer Perceptron (MLP) for analysis and perdition of stock market. We have used Bombay Stock Exchange (BSE) data set to analysis of stock market prediction.**

Keywords: Stock Price Forecasting, Prediction, Data Mining.

I. INTRODUCTION

Stock prediction is very important key factors for investors in financial domain,. The accurate prediction of stock is very challenging task and helps to profit of investors. Data mining based predictive techniques play very important role to analysis and predict of stock price. Prediction is one of the important applications of data mining and using in various fields especially in financial domain. Many researchers [3] attempts to predict stock prices by applying statistical and charting approaches, but those methods lacks behind heavily due to human biased decisions on stock market based on day to day mind set of human behaviour. Data mining is a suitable way to find out hidden patterns which was not possible by traditional approaches[1].

There are various authors have worked to analysis of prediction of stocks. Upadhyay A. et al (2012) [10] performed a study on the Multi Logistic Regression Model to determine the factors which significantly affect the performance of the company in the stock market. The relation between financial ratios and stock performance of the firms has been analyzed with the help of binary logistic regression. Kalmegh S. et al (2015) [9] discussed REP Tree, Simple cart, and Random Tree classification algorithms on Indian news domain. As a result it is found that Random Tree algorithm performs best in categorizing all the news. Patil S. S. et al. (2016) [5] performed a study on SVM algorithm and work on the large dataset value which collected from different global financial markets. Also SVM does not give a problem of over fitting. Correlation analysis indicates strong interconnection between the Market stock index and global markets that close right before or at the very beginning of trading time. Nair B.B. et al. (2010) [6] suggested the design and performance evaluation of hybrid decision tree- rough set based system for predicting the next day's trend in the Bombay Stock Exchange (BSESENSEX).The proposed hybrid decision tree-rough set based trend prediction system produces better performance.

II. ARCHITECTURE OF PROPOSED SYSTEM

The architecture of proposed system is depicted in Figure1. Proposed research work introduces a framework to predict the models using BSE data set. In this frameworks dataset is given to pre processing stage which further predicted by selected classifier. Machine learning tools WEKA[2] is used to analyze the performance of models.

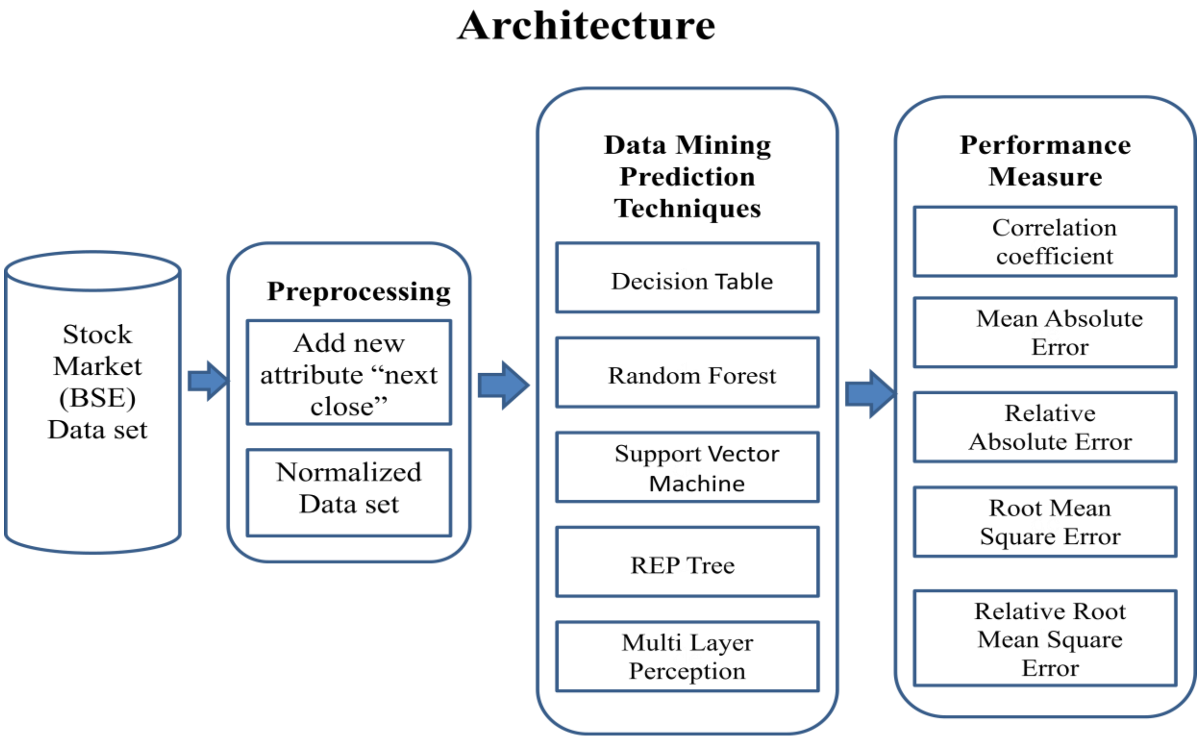

Figure 1 : External Architecure of Framework

A. *BSE Data Set*

BSE data set is collected from Bombay Stock Exchange of site (web source http://www.bseindia.com/indices/IndexArchiveData.aspx) [11] to analysis of data and predict the stock. The datasets are numeric dataset from 2 Jan 2012 to 2 Jan 2017 which has 1240 instances , 4 features namely open, high, low and close and 1 class level that is next day close with different continuous value. The detail of data set is shown in table 1.

Table 1: Data Set Description

Data Set Name	No .of Instances	Raw Attributes & Its Description
Stock Market Data Set(BSE Data Set)	1240	1. Open - The first traded price during the day or in the morning 2. High - The highest traded price during the day. 3. Low - The lowest price traded during the day. 4. Close - The last price traded during the day.

B. *Predictive Model*

In this research work, we have used various predictive models for prediction of stock. The predictive models are analyzed using Correlation Coefficient (CC), Mean Absolute Error (MAE), Relative Absolute Error (RAE), Root Mean Square Error (RMSE) and Relative Root Mean Square Error (RRMSE).

C. *Pre-processing*

To achieve the better performance, pre-processing is important step of data mining. In paper we have added the next close field as class level and normalize the data.

D. *Methods*

We have used different prediction techniques such as Decision Table, Random Forest, Support Vector Machine, REP Tree, Multi Layer Perceptron. Description of all these methods are as follows:

➢ ***Decision Table(DT)***

Decision Tables [4] are classification models elicited by machine learning algorithms and are used for creating predictions. A decision table consists of a hierarchical table within which entry in a higher level table gets broken down by the values of a pair of additional attributes to make another table.

➢ ***Random Forest(RF)***

Random Forest (Source: http://www.listendata.com/search?q=random+forest) is one of the most widely used machine learning algorithm for classification. It can also be used for regression model (i.e. continuous target variable) but it mainly performs well on classification model (i.e. categorical target variable). It has become a lethal weapon of modern data scientists to refine the predictive model. The best part of the algorithm is that there are a very few assumptions attached to it so data preparation is less challenging and results to time saving. Random forest/decision tree, classification model refers to factor/categorical dependent variable and regression model refers to numeric or continuous dependent variable.

- ***REP Tree(RT)***

Rep Tree uses [9] the regression tree logic and creates multiple trees in different iterations. After that it selects best one from all generated trees. That will be considered as the representative. In pruning the tree the measure used is the mean square error on the predictions made by the tree.

- ***Support Vector Machine (SVM)***

SVM [5] uses linear model to implement nonlinear class boundaries through some nonlinear mapping the input vectors x into the high-dimensional feature space. A linear model constructed in the new space can represent a nonlinear decision boundary in the original space. In the new space, an optimal separating hyper plane is constructed. The points on either side of the separating hyperplane have distances to the hyperplane. The smallest distance is called the margin of separation.

- ***Multilayer Perceptron (MLP)***

Multilayer Perceptron (MLP) network models [7] are the popular network architectures used in most of the research applications in medicine, engineering, mathematical modelling, etc..In MLP, the weighted sum of the inputs and bias term are passed to activation level through a transfer function to produce the output, and the units are arranged in a layered feed-forward topology called Feed Forward Neural Network (FFNN).

E. Performance Measures

To check the robustness of model we have calculated the various performance measures like Correlation Coefficient (CC), Mean Absolute Error (MAE), Relative Absolute Error (RAE), Root Mean Square Error (RMSE) and Relative Root Mean Square Error (RRMSE). Table 2 shows that various performance measures.

Table 2: Performance measures

Error Measures	**Formula**		**Description**
Correlation Coefficient	* Karl Pearson Coefficient		The correlation coefficient represent the degree of relational strength between two variables.
Absolute Error	Actual Value-Predicted Value		Absolute error is a measure of how far 'off' a measurement is from a true value.
	MAE $\sum(\vert f(xi)-yi\vert)/N$	RMSE $\sqrt{}\sum(f(xi)-yi\vert)^2/N$	
Relative Error	Absolute Error / Actual Value		Relative error expresses how large the absolute error is compared with the total size of the object you are measuring.
	RAE $\sum(\vert f(xi)-yi\vert)/\sum\vert f(xi\vert)$	RRSE $\sqrt{}\sum(f(xi)-yi)2/\sum(\bar{y}i)-yi)2$	

III. RESULTS AND DISCUSSION

This experiment is done in WEKA data mining software in window environment. We have used 10-fold cross validation to divide the dataset into 10 partitions where 1 part is used to test the model and rest of partition is used to train the model. We have applied the data set into different predictive models to analysis of stock prediction as shown in table 3. The robustness of model check in the terms of CC, MAE, RMSE, RAE and RRAE. The four measures are error measures like MAE, RMSE, RAE and RRAE. Table 3 shows that performance measures of different models where maximum value of Correlation Coefficient produced by the Support Vector Machine. The Correlation Coefficient measure basically describes the relational strength between actual and predicted value. High value of this measure indicates the good prediction capability of the model. The next category of performance measure are concern to the absolute error as its name implies this error measure simply shows a magnitude difference between actual and predictive values irrespective to the size of actual value, so in this category MAE and RMSE has been taken and SVM again produces a less error as compare to other models, which are MAE 0.0055, RMSE 0.0075. Next category of performance measure belongs to the relative error which is evaluated with respect to the size of actual value of data, again in this category SVM is leading as compare to other models and produces less error values which are RAE 4.30%., RRSE 5.38%. Finally we recommended SVM is robust and better predictive model for prediction of stock price.

Table 3: Performance measures of various predictive models

Predictive Model	**CC**	**MAE**	**RMSE**	**RAE**	**RRSE**
DT	0.9949	0.0114	0.014	8.90%	10.11
RT	0.998	0.0065	0.0087	5.064%	6.2843
RF	0.9983	0.0061	0.0081	4.72%	5.82
SVM	**0.9986**	**0.0055**	**0.0075**	**4.30%**	**5.38**
MLP	0.9962	0.0085	0.0121	6.63%	8.76

IV. CONCLUSION

Data mining based predictive models are very beneficial to analysis of model in different domain. In financial domain, predictive models play very important role for stock price prediction and important key factor for investors. We have applied the BSE data set to the different predictive model to check the robustness and compare the performance of models in terms of CC, MAE, RMSE, RAE and RRAE. Among all these models, the SVM produces better result as compare to others. The SVM model produces 0.9986, 0.0055, 0.0075,,4.30% and 5.38 of CC, MAE and RMSE RAE and RRSE respectively.

REFERENCES

1. J. han , M. Kamber," Data Mining Concepts and Techniques, published by Morgan Kauffman,2nd ed.2, 2006.
2. Web source: http:// www.cs.waikato.ac.nz/~ml/weka/ last accessed on Jan. 2017.
3. S. Prasanna and D. Ezhilmaran," An analysis on Stock Market Prediction using Data Mining Techniques", International Journal of Computer Science & Engineering Technology, Vol. 4 No. 02, pp. 49-51 , 2013
4. S.M. Mythili and R.M. Mohamed Shanavas," An Analysis of students' performance using classification algorithms", IOSR Journal of Computer Engineering, Vol. 16, Issue 1, pp. 63-69, 2014.
5. S.S. Patil, K Patidar. And M. Jain. " Stock Market Prediction Using Support Vector Machine", International Journal of Current Trends in Engineering & Technology, Vol. 02, Issue: 01, pp. 18-25 ,2016.
6. B. B. Nair, P.V. Mohandas. and R.N. Sakthivel ," A Decision Tree- Rough Set Hybrid System for Stock Market Trend Prediction", International Journal of Computer Applications,Vol. 6 No. 9, pp. ,2010.
7. P. Venkateshan and S. Anitha, ,"Application of a radial basis function neural network for diagnosis of diabetes mellitus". Current Science, Vol. 91, NO. 9, pp. 1195-1199 , 2006.
8. Source: http://www.listendata.com/search
   ?q=random+forest (accessing date : march 2017).
9. S. Kalmegh, " Analysis of WEKA Data Mining Algorithm REP Tree, Simple Cart and Random Tree for Classification of Indian News", International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 2, pp. 438-446, 2015.
10. A. Upadhyay, G. Bandyopadhyay, and A. Dutta , " Forecasting Stock Performance in Indian Market using Multinomial Logistic Regression", Journal of Business Studies Quarterly 2012, Vol. 3, No. 3, pp. 16-39, 2012.
11. web source http://www.bseindia.com/indices/IndexArchiveData.aspx) (browsing date : March 2017).