Prediction of Stock Market using Ensemble Model

B. Narayanan
Assistant Professor
Computer Science and Engineering Wing
D.D.E, Annamalai University

M. Govindarajan
Assistant Professor
Department of Computer Science and Engineering
Annamalai University

ABSTRACT
In the modern Digital Era, Data Mining is the powerful area for analyzing the large data sets to get unexpected relationships (models). The analysis of statistical data on sequential data points measured at regular time interval over a period of time is time series analysis. Time series analysis is used in predicting future occurrence of a time based event. One of the main areas where time series analysis is implied is in stock market prediction. The two important classification ways are Support Vector Machine (SVM) and Naïve Bayes. SVM is a method used for the foreseeing of financial time based data sets. It uses a function called risk which contains a mistake and a term. The basic principle which is used to obtain this may be called as minimization of structural risk. Naïve Bayes model assigns class labels for problem instances which can be denoted as vectors of feature measurements. For a given group variable, it takes into consideration that the numerical record of a specified feature is unique of others. Purpose of the present investigation is to develop an ensemble model namely AdaSVM and AdaNaive to analyse the stock data by comparing SVM and Naïve Bayes methods. The performance evaluation measures such as accuracy and classification error were computed individually for the stock market data set. The experimental result shows AdaSVM and AdaNaive is acceptable than SVM and Naïve Bayes.

General Terms
Stock market prediction, Time Series Analysis.

Keywords
Data Mining, Time Series, Stock market, Support Vector Machine, Naïve Bayes, AdaSVM, AdaNaive, Accuracy, Classification Error.

1. INTRODUCTION
Data mining is a specialized area in computer science in which one can able to extract the information from the large data base so as to make that information to usable structure for later use. The machine learning is the sub area of data mining where a model is developed in the computer by learning concept. This means that the model learns by training and testing over the given data. The model finally predicts new instances of data by making use of learning concept. In this paper, classifier techniques namely Support Vector machines and Naïve Bayes commonly called as predictive data mining techniques for analyzing time based data sets are used. The prediction which uses classifier based techniques gives better results [1]. The process of joining weak or inappropriate prediction rules for creating a machine based learning method to increase the prediction accuracy and to minimize the error rate is known as boosting.

A country's capital market largely depends on the depth and width of the stock base. The financial growth of a country largely depends on the expansion and development of long term capital. Thanks to corporate world, the scope of capital gets widened all over the world. The great renaissance noticed in the industrial world all over the world is due to shares and stocks. Not only big companies and investors are widening but small investors, individuals, salaried people, fixed income group are also nowadays interested in the purchase and sale of shares. Shares are purchased when the prices are low in the market and sold when they are high. The margin is the profit to the investor. Small investors are now busy in operating their DEMAT account for purchase and sale of shares. Security and Exchange Board of India (SEBI), The Central Bank, the RBI in India keep a close watch and regulate Stock exchanges. The value of share price is highly volatile and is subjected to fluctuations from time to time. The time series analysis is a much needed technology in the present situation, which is used to predict the futuristic results. A minor improvement in the performance of stock market prediction can give a great profit.The recorded data x_{ts} in which every item is recorded at a particular point of time 'ts' is called as time series. Two classes of time series such as discrete-time and continuous-time are available [2]. In former, the data are noted at specified point of time and later the recordings are noted continuously over a period.

The work is presented as follows. The related works carried out previously are discussed in section II. Materials and Methods used for the study are explained in section III. Proposed methodology is given in section IV. Performance evaluation measures taken into account for the analysis are described in section V. Experimental results and discussion using the proposed method is depicted in section VI. In fine, conclusions and future work is mapped in section VII.

2. RELATED WORKS
Numbers of reputed computer science researchers in the field of forecasting have always compared different similar works with their findings. Some of them are given here under.Wei Huang et al [3] compared the forecasting ability of Support Vector Machine (SVM) with those of Back propagation Neural Networks, Quadratic Discriminant Analysis and Linear Discriminant analysis. It is observed from the study results that SVM is better than other classification methods.

Mohamed M. Mostafa [4] uses Multi-Layer Perceptron (MLP) neural networks as well as regression neural networks for forecast the KSE (Kuwait Stock Exchange) closing price. The data set is used for the period 2001-2003. It is observed from the experimental results that the neuro- computational model performs better than the traditional statistical techniques.Hyun Joon Jung et al [5] have proposed a Binary Stock Event Model (BSEM) to predict future trends of the stock market using selected features. The authors have used two techniques as Bayesian Naive Classifier and Support Vector Machine. The prediction accuracies are demonstrated around 70-80% in a day's experiment.Jigar Patel et al [6] uses two kinds of input data such as stock trading data (technical parameters) and trend deterministic data for those technical parameters. The work distinguishes different

models viz., Artificial Neural Network (ANN), Support Vector Machine (SVM), random forest and Naïve Bayes. It is observed from the study that for first input data, random forest performs better than others. The results for the said techniques have improved for the subsequent data.

Shin-Fu Wu, Shie-Jue Lee [7] opines that two kinds of modeling options for developing forecasting models can be applied. The Global modeling is different from the queries of the user. For each query, the local model builds a native model. The study uses local modeling strategy to find out the acceptability of applying local model with Neural Network (NN), Adaptive Neuro-Fuzzy Inference System (ANFIS), and Least Squares Support Vector Machine (LS-SVM) for time data forecasting. It is observed from the study results that local modeling improves the performance. Michel Ballings et al [8] compare the ensemble methods such as Random Forest, AdaBoost and Kernel Factory against the single classifier models namely Neural Networks, Logistic Regression, Support Vector Machines and K-Nearest Neighbor. The data from 5767 European companies are taken for the study. The characteristic curve is used as the performance measures. The study results indicate that the Random forest is much superior algorithm.The researchers of this paper, in the same way have embarked to compare their proposed models AdaSVM and AdaNaive with SVM and NaiveBayes.

3. MATERIALS AND METHODS USED

3.1 Data Set

The Historical Time Series data collected from www.datamarket.com is used for the study. The data contains records of stock values of various companies. The CSV pattern possesses a record for each line of text akin to the data for a particular day. The record attributes are placed as Date, cname, Start, Max, Min, End, and Quantity for any point of day. The 'Date' represent the day in which the Stock price is taken into account. The 'Start' represents the price value of the stock at the opening moments of a day. The 'Max' represents the highest price at which a particular stock is sold out on the day. The 'Min' represents the lowest price at which a particular stock is sold out on the day. The 'End' represents the closing price of the stock in the end of the day. The 'Quantity' represents the number of stocks that have been sold out on the particular day. The name of the company whose stock is being analyzed is represented by the attribute 'cname' which is not taken for the study.

The dataset is separated for training and testing. The training dataset consists of 3777 records and testing dataset consists of 2500 records. Training dataset is used for generating models from the classification algorithm. To test the accuracy of the generated model it is applied on the testing dataset..

3.2 Machine Learning Methods

3.2.1 Support Vector Machine

The machine learning technique that is used in the pattern recognition area is SVM. The SVM is used to minimize the structural risk. Nearness of data points in the decision surface (hyperplane) indicates support vectors. The object of SVM is to make a model based on the training data. It gives the target values lent by the attributes of the test data. The SVM obtain a separating hyperplane which divides the data with widest margin for linearly separable data. The SVM maps linearly inseparable data in the input space into big dimensional space by $x \in \mathbb{R}^I \rightarrow \phi(x) \in \mathbb{R}^H$ where $\phi(x)$ is the kernel function used to obtain the separating hyperplane. Any symmetric function which satisfies the Mercer's condition is known as Kernel function (Courant and Hilbert, 1953). The fundamental kernel used in SVM are linear, polynomial, radial basis function (RBF) and sigmoid. The application of the support vector machine and its fusion techniques for the analysis of time series data gives promising results [9, 10].

3.2.2 Naïve Bayes

The Bayesian theorem is the principle behind naïve bayes classification technique. This technique is very much suitable when the value of the inputs is very high. Bayes classifiers are also called as Simple Bayes or Idiot Bayes. The Bayes theorem is $p(a_c|b) = p(b|a_c)p(a_c)/p(b)$ where $p(a_c|b)$ denotes probability of b instance in class a_c . For the given class a_c, $p(b|a_c)$ represents the probability of producing b instance. $p(a_c)$ is probability of the occurrence of a_c and $p(d)$ is probability of the occurrence of d. This theorem of classification is easily done when the attribute is single. The classification is done by Naïve Bayes by extending the above theorem for more than one attributes by using the formula $p(b|a_c) = p(b1|a_c) * p(b2|a_c) * \ldots * p(bn|a_c)$ where $p(b1|a_c)$ represents the probability of class a_c which produces the value for attribute 1, $p(b2|a_c)$ indicates the probability of class a_c which develops the value for 2nd attribute. The calculation is done by Naïve Bayes using the formula by the assumption that attributes must have independent distributions. The Naïve bayes is used in financial forecasting to get better results [11].

3.3 Ensemble Model

3.3.1 AdaBoost

The ability of the ensemble model is to expand the classification accuracy by creating more than one classifier. The ensemble model supports decision making by joining the results of their classification techniques. The accuracy of the given algorithm is improved by boosting method. The AdaBoost is one such boosting algorithm developed by Yoav Freund and Robert E. Schapire in 1995 [12]. The input for training set of AdaBoost algorithm is $(A_1, B_1), (A_2, B_2), \ldots, (A_m, B_m)$ where A_i represents space set A and B_i represents space set B. Assumption is made such that $B = (-1, +1)$. For repeated sets $Z = 1, \ldots Z$ AdaBoost calls the base or weak algorithm. The algorithm pertain weights on the training set. In starting point, the weights are evenly distributed. But for the subsequent training, the weights are increased for the not properly classified example. By doing so, the weak learner must be trained on hard examples. This process will improve the accuracy of classification.

4. PROPOSED METHODOLOGY

In this work, the motivation is to analyze the time based data set and to forecast the stock market price more precisely than the existing models. Therefore the ensemble models such as AdaSVM and AdaNaive are developed by using AdaBoost technique. The ability of the AdaBoost technique is that it focuses on the weak learners that are hard to learn. SVM when combined with AdaBoost (AdaSVM) will classify efficiently by providing weak learners with proper training. The same approach is used for Naïve Bayes classifier, in which AdaBoost based Naïve Bayes (AdaNaive) is used to better classify the data. The proposed work is organized as shown in Figure 4.1.

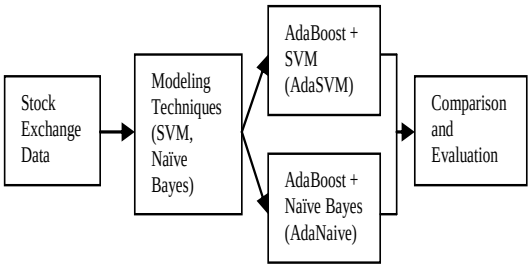

Fig 4.1: Overview of the proposed work

5. PERFORMANCE EVALUATION MEASURES

5.1 Accuracy

The closeness to a measured value or the standard set is called accuracy. In time series analysis, the forecasting value which is nearest to the actual value is taken as accuracy. The formula for accuracy is $A = (TP + TN)/(TP + FP + FN + TN)$ where the true positive cases are denoted by TP, true negative cases are denoted by TN, FP and FN are denoted for false positive cases and false negative cases respectively.

5.2 Classification Error

The classification Error (E) of any technique 't' are the cases not correctly classified $(FP + FN)$. The formula for calculating classification Error is $E_t = \left(\frac{F}{N}\right) * 100$ where t represents the technique, F denotes number of items classified incorrectly and N reveals total number of samples.

6. EXPERIMENTAL RESULTS AND DISCUSSION

In this paper, time series data forecast is done by the classification techniques SVM, AdaSVM, Naïve Bayes, AdaNaive. The Data set is seperated into couple of groups, one for training and another for testing the classification algorithms. Rapidminer data analysis tool is used to implement the classification algorithms. Secondary data kept in CSV file is first loaded using the "Read CSV" operator in rapidminer tool. From the loaded data only a subset of data is selected for implementing classification process. The operator "Select Attributes" is used to select a subset from the original data. The selected subset is processed with the "X-Validation" operator which first builds the classification model and then the model is validated using the test data. In "X-Validation" operator the classification functions AdaBoost based SVM (AdaSVM), SVM, AdaBoost based Naïve Bayes (AdaNaïve) and Naïve Bayes are applied. Performance operator is used to evaluate the performance of the classification algorithm. The analysis of result shows that the proposed AdaBoost Support Vector Machine and Naïve Bayes are shown to be superior to individual approaches for stock market prediction problem in terms of classification accuracy and error. Performance obtained for both the classification algorithms are given in Table 6.1.The accuracy and Classification Error of prediction for the specified machine learning techniques is shown in the Fig 6.1. The proposed combined models show significantly improvement of classification accuracy and lower error than the base classifiers. The results show that when combined with AdaBoost both SVM and Naïve Bayes algorithm have better performance values than compared to the SVM and Naïve Bayes algorithm.

Table 6.1. Performance evaluation of Existing and proposed (AdaSVM & AdaNaive) techniques

Measures	Existing SVM	Proposed AdaSVM	Existing Naïve Bayes	Proposed AdaNaïve
Accuracy	93.86%	94.33%	88.32%	97.19%
Classification Error	6.14%	5.67%	11.68%	2.81%

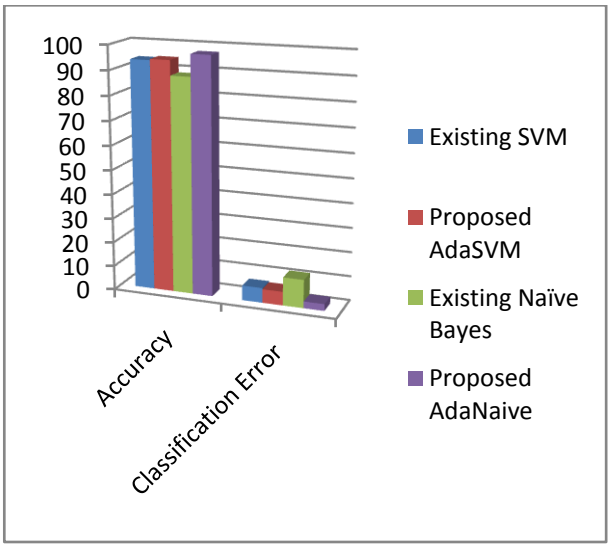

Fig 6.1: Accuracy and Classification Error for stock prediction of the specified techniques (in %)

7. CONCLUSION AND FUTURE WORK

In this study, the time series analysis for stock market prediction is carried out by two existing techniques *viz.*, Support Vector Machine and Naïve Bayes. The results indicate that the accuracy, classification error of prediction by SVM is 93.86% and 6.14% respectively where as Naïve Bayes is 88.32% and 11.68% respectively. For the same set of input data, the proposed AdaSVM produces 94.33% of accuracy and 5.67% of classification error where as the proposed AdaNaive produces 97.19% of accuracy and 2.81% of classification error. The results indicate that there is substantial increase in the accuracy of prediction and amicable decrease in the percentage of classification error by both the proposed techniques. In future, the study to identify whether changing the technique reflects the result or by increasing the input data set for the same technique results change in the findings can be extended. Importance of predictions in stock market cannot be over emphasized. Continuous research for landing on most acceptable method of forecast is needed. This paper is an early step towards the end goal.

8. REFERENCES

[1] Itzamá López-Yáñez, Leonid Sheremetov, Cornelio Yáñez-Márquez, "A novel associative model for time series data mining**",** Pattern Recognition Letters, Volume 41, Pages 23-33, 2014.

[2] Peter J. Brockwell, Richard A. Davis, "Introduction to Time Series and Forecasting", Springer, 2002.

[3] Wei Huang, Yoshiteru Nakamori, Shou-Yang Wang, "Forecasting stock market movement direction with support vector machine", Computers & Operations Research, Volume 32, Issue 10, Pages 2513-2522, 2005.

[4] Mohamed M. Mostafa, "Forecasting stock exchange movements using neural networks: Empirical evidence from Kuwait", Expert Systems with Applications, Volume 37, Issue 9, Pages 6302-6309, 2010.

[5] Hyun Joon Jung, Aggarwal J.K, "A Binary Stock Event Model for stock trends forecasting: Forecasting stock trends via a simple and accurate approach with machine learning", 11th International Conference on Intelligent Systems Design and Applications (ISDA), Pages 714-719, 2011.

[6] Jigar Patel, Sahil Shah, Priyank Thakkar ,K Kotecha ,"Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques", Expert Systems with Applications, Volume 42, Issue 1, Pages 259-268, 2015.

[7] Shin-Fu Wu, Shie-Jue Lee, "Employing local modeling in machine learning based methods for time-series prediction", Expert Systems with Applications, Volume 42, Issue 1, Pages 341-354, 2015.

[8] Michel Ballings, Dirk Van den Poel, Nathalie Hespeels, Ruben Gryp, "Evaluating multiple classifiers for stock price direction prediction", Expert Systems with Applications, Volume 42, Issue 20, Pages 7046-7056, 2015.

[9] Kamil Żbikowski, " Using Volume Weighted Support Vector Machines with walk forward testing and feature selection for the purpose of creating stock trading strategy", Expert Systems with Applications, Volume 42, Issue 4, Pages 1797-1805, 2015.

[10] Shuhaida Ismail, Ani Shabri, Ruhaidah Samsudin, "A hybrid model of self-organizing maps (SOM) and least square support vector machine (LSSVM) for time-series forecasting", Expert Systems with Applications, Volume 38, Issue 8, Pages 10574-10578, 2011.

[11] Kurniady, A., Kosala, R., "Knowledge-based integrated financial forecasting system", International Conference on Computer Research and Development (ICCRD), Volume 1, Pages 120-124, 2011.

[12] Yoav Freund and Robert E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting", Journal of Computer and System Sciences, Volume 55, Issue 1, Pages 119–139, 1997.