IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 3, Ver. II (May-Jun. 2014), PP 06-10
www.iosrjournals.org

Application of Machine Learning Algorithm in Indian Stock Market Data

[1]Sanjaya kumar Sen , [2]Dr.Subhendu kumar Pani
[1]*Associate Prof., Dept. Of CSE,OEC,BBSR,Odisha*
[2]*Associate Prof., Dept. Of CSE,OEC,BBSR,Odisha*

Abstract: *Prediction of Indian stock market data with data mining technique is one of the fascinating issues for researchers over the past decade. Statistical and traditional methods are no longer feasible for proper analysis of huge amount of data. With the help of Data mining technique information technology tool , it is able to uncover hidden patterns and predict future trends and behavior in stock market. In this paper there are combination of four supervised machine learning algorithms, classification and regression tree (CART), linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) are proposed for classification of Indian stock market data. These resulted forms help market analyst to make decision on selling, purchasing or holding stock for a particular company in Indian stock market. In section IV and V, experimental results and performance comparison section show that classification and regression tree misclassification rate is only 56.11% whereas LDA and QDA show 74.26% and 76.57% respectively. Smaller misclassification reveals that CART algorithm performs better classification of Indian stock market data as compared to LDA and QDA algorithms.*

Keyword: *Machine learning algorithm; classification; patterns*

I. Introduction

Knowledge discovery in databases denotes the complex process of identifying valid, novel, potentially useful and ultimately understandable patterns in data Data *mining* refers to a particular step in the KDD process. According to the most recent and broad definition [1], "data mining consists of particular algorithms (methods) that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns (models) over the data.

In this paper classification of Indian stock market data is done using machine learning algorithm which are under supervised machine learning . With the help of machine learning algorithm, market analyst are able to make decision on selling, purchasing or holding stock for a particular company in Indian stock market for profit on selling, purchasing or holding stock for a particular company in Indian stock market for profit.The popular data mining algorithms can also be used in predictions of software projects[8]. The work proposed in this paper is unique as compared to other works in literature because we have used a combination of supervised machine learning algorithms for classification of Indian stock market data while other works consist of unsupervised machine learning algorithms.

II. Machine Learning Algorithms

In this paper three supervised machine learning algorithms CART (classification and regression tree), LDA (linear discriminant analysis) and QDA (quadratic discriminant analysis) are used for classifying Indian stock market data.

A. Classification and Regression Tree (CART)

Classification and Regression Trees is a classification method which uses historical data to construct so-called decision trees[2]. Decision trees are then used to classify new data. In order to use CART we need to know number of classes a priori. CART methodology was developed in 80s by Breiman, Freidman, Olshen, Stone in their paper "Classification and Regression Trees" (1984). For building decision trees, CART uses so-called learning sample - a set of historical data with pre-assigned classes for all observations. Decision trees are represented by a set of questions which splits the learning sample into smaller and smaller parts CART algorithm will search for all possible variables and all possible values in order to find the best split – the question that splits the data into two parts with maximum homogeneity. The process is then repeated for each of the resulting data fragments. Here is an example of simple classification tree, used by San Diego Medical Centre for classification of their patients to different levels of risk. It can deal with both numeric and categorical attributes and can also handle missing attributes [3] The CART monograph focuses on the Gini rule, which is similar to the better known entropy or information-gain criterion [4]. For a binary (0/1) target the "Gini measure of impurity" of a node t is:

Classification and regression tree always calculates class frequencies in any node relative to the class frequencies in the root. For a binary (0/1) target any node is classified as class 1 if, and only if, Classification and regression tree provide automatic construction of new features within each node and for the binary target. Single feature is given by following equation: Where, is the original attribute vector and is a scaled difference of means vector across the two classes [4].

B. Adaboost

The AdaBoost algorithm, introduced in 1995 by Freund and Schapire [5], solved many of the practical difficulties of the earlier boosting algorithms, and is the focus of this paper. AdaBoost, short for Adaptive boosting is a machine algorithm, formulated by Yeave Freud and Robert Scapire. eight assigned toht Schapire and Singer [6] show how AdaBoost and its analysis can be extended to handle weak hypotheses which output real-valued or confidence-rated predictions It is a meta-algorithm and can be used in conjunction with many other learning algorithms to improve their performance. AdaBoost can be viewed as a form of functional gradient descent, as observed by Mason et al. [7] and Friedman AdaBoost is an algorithm for constructing a " strong" classifier as linear combination. AdaBoost is adaptive in the sense that subsequent classifiers built are tweaked in favour of those instances misclassified by previous classifiers. AdaBoost is sensitive to noisy data and outliers. In some problems, however, it can be less susceptible to the over fitting problem than most learning algorithms. The classifiers it uses can be weak (i.e., display a substantial error rate), but as long as their performance is not random (resulting in an error rate of 0.5 for binary classification), they will improve the final model. Even classifiers with an error rate higher than would be expected from a random classifier will be useful, since they will have negative coefficients in the final linear combination of classifiers and hence behave like their inverses.

AdaBoost is an algorithm for constructing a "strong" classifier as linear combination
$f(x) = \sum \alpha t ht(x)$ of "simple" "weak" classifiers ht(x).
Terminology

ht(x) ... "weak" or basis classifier, hypothesis, "feature"

H(x) = sign(f(x)) ... ''strong" or final classifier /hypothesis

C. Bagging

The way of combining the decisions of different models means amalgamating the various outputs into a single prediction. The way of doing to do this is to calculate the average. In bagging the models receives equal weights. In case of bagging suppose that several training datasets of the same size are chosen at random from the problem domain. This is radar data gathered by the Space Physics Group at Johns Hopkins University (see Sigillito et. al. [9]) Suppose using a particular machine learning technique to build a decision dtree for each dataset, we might expect these trees to be practically identically and to make the same prediction for each new test instance. This is a disturbing fact and seems to cast a shadow over the whole enterprise. In bagging the models receive equal weights, whereas in boosting weighting is used to give more influence to the more successful one just as an executive might place different values on the advice of different experts depending on how experienced they are. Buntine [10] gave a Bayesian approach, Kwok and Carter [11] used voting over multiple trees generated by using alternative splits, and Heath used voting over multiple trees generated by alternative oblique splits. Dieterrich [11] showed that a method for coding .An Introduction to the Bootstrap". Chapman and Hall.any class problems into a large number of two class problems increases accuracy.To introduce bagging, several training datasets of the same size are chosen at random from the problem domain. Suppose using a particular machine learning technique to build a decision tree for each dataset, we might expect these trees to be practically identically and to make the same prediction for each new test instance.

D. Logitboost

LogitBoost is a boosting algorithm formulated by Jerome Friedmome, Trevor Hastie, and Robert Tibshirani. The original paper casts the Adaboost algorithm into a statistical framework. Specifically, if one considers AdaBoost as a generalized additive model and then applies the cost functional of logistic regression one can derive the LogitBoost algorithm.

E. Grading

We investigate another technique, which we call *grading*. The basic idea is to learn to predict for each of the original learning algorithms whether its prediction for a particular example is correct or not. We therefore train one classifier for each of the original learning algorithms on a training set that consists of the original examples with class labels that encode whether the prediction of this learner was correct on this particular

example. The algorithm may also be viewed as an attempt to extend the work of Bay and Pazzani (2000)[13] who propose to use a meta-classification scheme for characterizing model errors. Hence in contrast to stacking—we leave the original examples unchanged, but instead modify the class labels. The algorithm may also be viewed as an attempt to extend the work of Bay and Pazzani (2000)[13] who propose to use a meta-classification scheme for characterizing model errors. Their suggestion is to learn a comprehensible theory that describes the regions of errors of a given classifier. While the step of constructing the training set for the meta classifier is basically the same as in our approach,1 their approach is restricted to learning descriptive characterizations but cannot be directly used for improving classifier performance. The reason is that negative feedback when the meta classifier predicts that the base classifier is wrong only rules out the class predicted by the base classifier, but does not help to choose among the remaining classes (except, of course, for two-class problems).

III. Performance Comparisons

In literature classification algorithm, classifier performance can be measured on the same data .This paper presents computational issues of three supervised machine learning algorithm i.e. classification and regression tree, linear and quadratic discriminant analysis algorithm on Indian stock market data for its classification with dedicated goal for maximizing profit of market analyst and investors to make decision for selling , purchasing or holding stock of a particular company on the basis of classification rule. Among three algorithms, classification and regression tree algorithm is good because result show that classification and regression tree algorithm classification results are easier to interpret and understand as compared to linear and quadratic discriminant analysis algorithm because it gives results in the form of tree structure. In order to compare the classification performance of three machine learning algorithm, classifiers are applied on same data and results are compared on the basis of misclassification and correct classification rate and according to experimental results in table 1, it can be conclude that classification and regression tree, supervised machine learning algorithm is best as compared to linear and quadratic discriminant analysis.

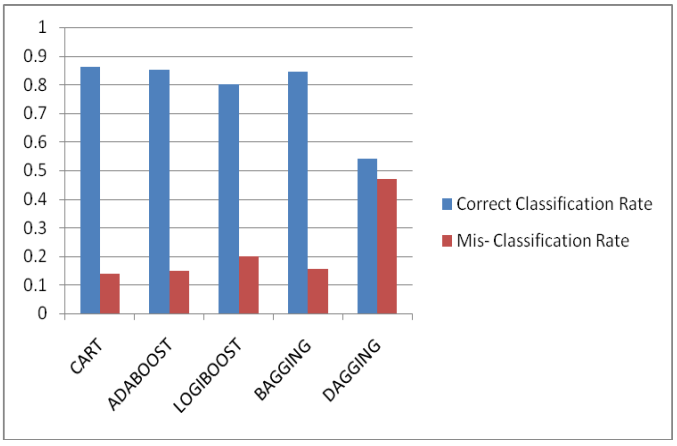
In literature classification algorithm, classifier performance can be measured on the same data. On the basis of results obtained CART algorithm is found better than other

Result Comparision Of Adaboost, Logiboost,Cart,Bagging & Dagging

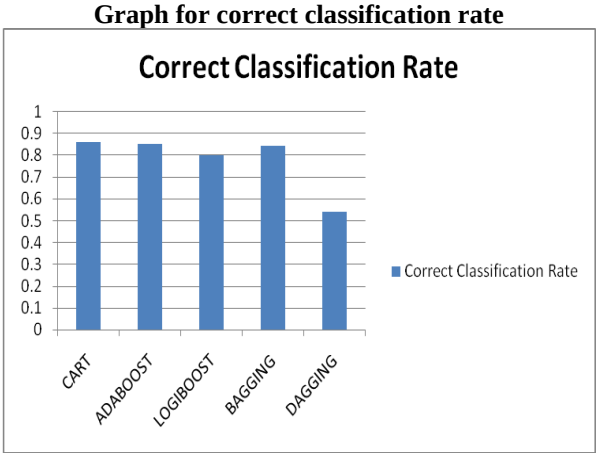
Algorithm Misclassification	Correct Classification Rate	Mis- Classification Rate
CART	**0.861**	**0.139**
ADABOOST	**0.852**	**0.148**
LOGIBOOST	**0.801**	**0.199**
BAGGING	**0.850**	**0.150**
DAGGING	**0.542**	**0.458**

According to obtained results of classification in table 1 following graph can be drawn.

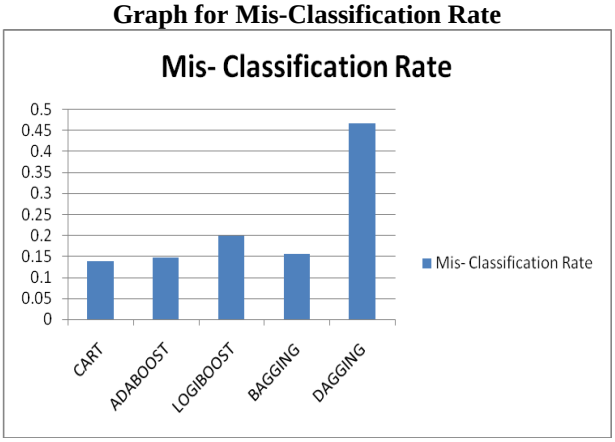
Graph for Correct Classification and Misclassification rate



Graph for correct classification rate



Graph for Mis-Classification Rate



IV. Conclusions

This paper represents computational issues of five supervised machine learning algorithms i.e. Adaboost algorithm, Logitboost algorithm, Classification and Regression technique algorithm, Bagging algorithm and Dagging algorithm for the analysis of stock on Indian stock market data for its classification and for the purpose of the maximizing profit of market analyst and investors to make and predict decision for selling , purchasing or holding stock of a particular company on the basis of classification rule. Among five algorithms, CART algorithm is good because result show that CART algorithm results are easier to interpret and understand as compared to Logiboost, Adaboost, Bagging and Dagging because it gives results in the form of tree structure. In order to compare the classification performance of three machine learning algorithm, classifiers are applied on same data and results are compared on the basis of misclassification and correct classification rate and according to experimental results in table 1. It is concluded that classification and regression tree, supervised machine learning algorithm is best as compared Logiboost, Adaboost, Bagging and Dagging on the basis of classification rule.

Reference

[1]. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. *Advances in Knowledge Discovery and Data Mining*. AAAI Press/MIT Press, Menlo Park, California/Cambridge, Massachusetts/ London, England, 1996.

[2]. Subhendu Kumar Pani and Amit Kumar and Maya Nayak," Performance Analysis of Data Classification Using Feature Selection" (October 24, 2013). The IUP Journal of Information Technology, Vol. IX, No. 2, June 2013, pp. 36-50

[3]. Matthew N. Anyanwu and Sajjan G. Shiva," Comparative Analysis of Serial Decision Tree Classification Algorithms," International Journal of Computer Science and Security, (IJCSS) Volume (3) : Issue (3) ,.pp: 230-240.

[4]. XindongWu ,Vipin Kumar , J. Ross Quinlan ,Joydeep Ghosh , Qiang Yang ,Hiroshi Motoda , et al "Top 10 algorithm in data mining according to the survey paper of Xindong wu et al know (inf syst (2008) @ springer verlog London limited 2007,pp:1-37, Dec. 2007.

[5]. 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, 55(1):119–139, August 1997

[6]. Robert E. Schapire and Yoram Singer. Improved boosting algorithms using confidence-rated predictions. In Proceedings of the Eleventh AnnualConference on Computational Learning Theory, pages 80–91, 1998. To appear, Machine Learni.

[7]. Mason, L., Baxter, J., Bartlett, P., Frean, M.: Functional gradient techniques for combining hypotheses. In: Advances in Large Margin Classifiers. MIT Press (2000).

[8]. Subhendu Kumar Pani and Satya Ranjan Biswal and Santosh Kumar Swain, A Data Mining Approach to Identify Key Factors for Systematic Reuse (October 31, 2012). The IUP Journal of Information Technology, Vol. VIII, No. 2, June 2012, pp. 24-34. Available at SSRN: http://ssrn.com/abstract=2169262

[9]. Sigillito, V. G., Wing, S. P., Hutton, L. V.,and Baker, K. B. () Classication of radar returns from the ionosphere using neural networks. Johns Hopkins APL Technical Digest, 10, 262-266.

[10]. Buntine, W. \Learning classication trees", Articial Intelligence Frontiers in Statistics, ed D.J. Hand, Chapman and Hall, London, 182-201.

[11]. Dietterich, T.G. and Bakiri, G. Error-correcting output codes: A general method for improving multiclass inductive learning programs, Proceedings of the Ninth National Conference on Articial Intelligence (AAAI-91),Anaheim, CA: AAAI Press.

[12]. Kwok, S., and Carter, C. (1990) Multiple decision trees, Uncertainty in Artificial Intelligence 4, ed. Shachter, R., Levitt, T., Kanal, L., and Lemmer,J., North-Holland, 327-33.

[13]. Bay, S. D., & Pazzani, M. J. (2000). Characterizing model errors and differences. In *Proceedings of the 17th International Conference on Machine Learning (ICML- 2000)*. Morgan Kaufmann.