IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Vol. 05, Issue 08 (August. 2015), ||V4|| PP 05-10

Application of Meta learning in Banking Sector

Sanjay Kumar Sen, Dr. B.K. Ratha.
Assistant Professor, Dept. of CSE . Orissa Engineering College, Bhubaneswar, Odisha, India
sanjaysen2k@gmail.com
Reader, Dept. of Computer Science and Application, Utkal University, Vani Vihar, Bhubaneswar.
vkramus@yahoo.com

Abstract: - The data mining technology is a process of identifying patterns and information from a huge quantity of data. In a single repository data base where data is stored in central site, then applying data mining algorithms on these data base, patterns are extracted which are converted into information. It becomes important area for financial organization like banking sector, insurance sector, stock market etc. Banking systems gather huge amount of data like customer information, transaction details, loan details, credit card details etc. It helps to find hidden patterns in a group and discover unknown relationship in the data. Instead of quantitative and statistical data characteristics, this technique helps in data interpretations for the banking sector. Data mining tools helpful for leading banking for credit scoring service and approval predicting payment lapse, fraud detecting etc. By applying data mining techniques bankers are getting advantage in competitive market. It is also helpful for the retailers for customer buying tradition and also desires. This proposed work in this paper is the combination of supervised machine learning algorithms, Classification and Regression Tree (CART), Adaboost and Logitboost, Bagging and Dagging are proposed for classification of banking data. These resulted forms help researchers to detect fraud in banking data. The experimental result shows the performance analysis of different meta-learning algorithms and also compared on the basis of misclassification and correct classification rate. Smaller misclassification reveals that bagging algorithm performs better classification of banking data fraud detection technique.. The proposed work in this paper is the combination of five supervised machine learning algorithms, Classification and Regression Tree (CART), Adaboost, Grading, Bagging are proposed for classification of banking data. The experimental result shows the performance analysis of different meta-learning algorithms and also compared on the basis of misclassification and correct classification rate. Smaller misclassification reveals that bagging algorithm performs better classification of banking data fraud detection technique.

I. INTRODUCTION

The modern technology has great influence on banking business. All traditional banks compelled to adopt and implement new technological approach not only to improve financial and customer relationship but also to reduce time and cost to survive in this competitive market. Some automated software which radically changed the basic definition of the definition of the banking business.

Data mining is the process of extracting hidden and unknown knowledge by applying statistical and machine-learning techniques for extracting interesting patterns from huge data base. Data mining is the application of statistical and machine-learning techniques for extracting interesting patterns from raw data [4]

The strong security systems maintained in banking industries to avoid fraud in internet banking. Though fraud in banking can not be removed drastically., but to the some extent can be minimized. Programs for Machine Learning, published in 1993 [1], that many applications of artificial intelligence are based on a model of knowledge that is usually employed by a human specialist. This paper helps to minimize the bank fraud by using different meta learning algorithm. Initial work on the forecasting system of banking system is done by Frankel and Rose using some indicators in relation to the emerging currency markets[2]. The first group indicates integrated indicators of interest rates of interest rates and output growth[3]. The second group includes national macroeconomic indicators like sudden changes the rate of production[3]. The third category includes indicators of price revaluation, the current account deficit and the level of debt. The fourth class of variables describes the determinants of a debt structure [10].

1.1Application of Data Mining in Banking Sector

Various areas where data mining can be used in financial sectors [5][6][7] like customer segmentation and profitability, credit analysis, predicting payment default, marketing, fraudulent transactions, ranking investments, optimizing stock portfolios, cash management and forecasting operations, high risk loan applicants, most profitable Credit Card Customers and Cross Selling. Certain examples where banking industry has been utilizing the data mining technology effectively as follows.

1.2Fraud Detection Fraud detection

[8][9] [10][11] [12] is the recognition of symptoms of fraud where no priorm suspicion or tendency to fraud exists. Fraud is defined as 'a deception deliberately practiced in order to secure unfair of unlawful gain. Fraud detection is the detection of criminal activities which occurs in commercial organizations such as banks, credit card issuing organizations, insurance agencies, mobile companies, stock market. Themmalicious users might be the actual customers of the organization or might be posing as a customer (also known as identity theft) [11]. Financial Organizations especially banking sectors follows mainly two approaches[5] [12]towards determining the fraud patterns, online transaction check and Offline transaction Check.

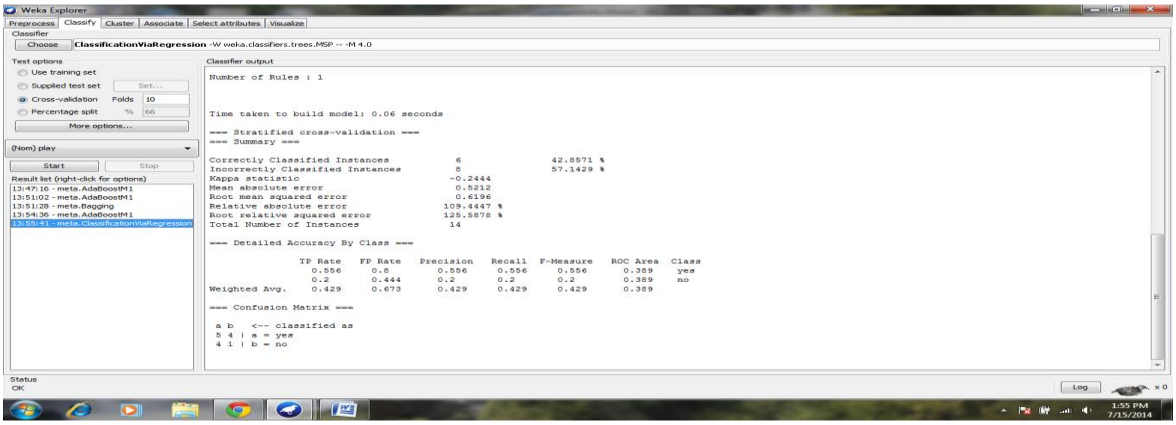
1.3 Risk Management

Data Mining is used to identify the risk factors in each department of banking business [6]. Credit Approval authorities in the financial organization used data mining techniques to determine the risk factors in lending decisions[13] by analyzing the data based on nationality, repayment capacity and so on. Retail marketing department uses data mining methodologies to find the reliability[14] and the behavior of credit card applicant[8] while selling the credit cards. They uses data mining techniques on existing customers to sell credit cards or increase customers credits or top up on credit card loans [15]. In commercial lending, data mining plays a vital role. In commercial lending, risk assessment is usually an attempt to quantify the risk of default or loss to the lender when making a particular lending decision or approving a credit facility [8].

Data Mining can be used to derive credit behavior [8] of individual borrowers with parameters card loans, mortgage value, repayment and using characteristics such as history of credit, employmen period and length of residency. A score is thus produced that allow a lender to evaluate the customer and decide whether the person is a good candidate for a loan, or if there is a tendency to become high risk of default[14] Customers who have been with bank for a longer periods of time, remained better with bank and have good credit history and have higher salaries/wages, are more likely to receive a loan than a new customer who has no credit history with the bank, or who earns low salaries/wages[17].Bank can reduce the risk factors to maintain a better position by knowing the chances of a customer to become default[18] [19].

II. ALGORITHM USED

1.Classification and Regression Tree(CART) It was introduce by Breimann 1984.It builds both classification and regression tree (Gini index measure is used for selecting splitting attribute. Pruning is done on training data set. It can deal with both numeric and categorical attributes and can also handle missing attributes. [12]. The CART monograph focus on the Gini rule which is similar to the better know entropy or information gain criterion [13]. For binary (0/1) target the 'Gini measure of impurity" of a node t is: Classification and regression free provide automatic construction of new features within each node and for the binary target[11].



Classifier: ClassificationViaRegression -W weka.classifiers.trees.M5P -- -M 4.0

Test options
○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) play

Result list (right-click for options)
13:47:16 - meta.AdaBoostM1
13:51:02 - meta.AdaBoostM1
13:51:28 - meta.Bagging
13:54:36 - meta.AdaBoostM1
13:55:41 - meta.ClassificationViaRegression

Classifier output

```
Number of Rules : 1



Time taken to build model: 0.06 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances           6               42.8571 %
Incorrectly Classified Instances         8               57.1429 %
Kappa statistic                         -0.2444
Mean absolute error                      0.5212
Root mean squared error                  0.6196
Relative absolute error                109.4447 %
Root relative squared error            125.5878 %
Total Number of Instances               14

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.556     0.8       0.556      0.556     0.556      0.389    yes
                 0.2       0.444     0.2        0.2       0.2        0.389    no
Weighted Avg.    0.429     0.673     0.429      0.429     0.429      0.389

=== Confusion Matrix ===

 a b   <-- classified as
 5 4 | a = yes
 4 1 | b = no
```

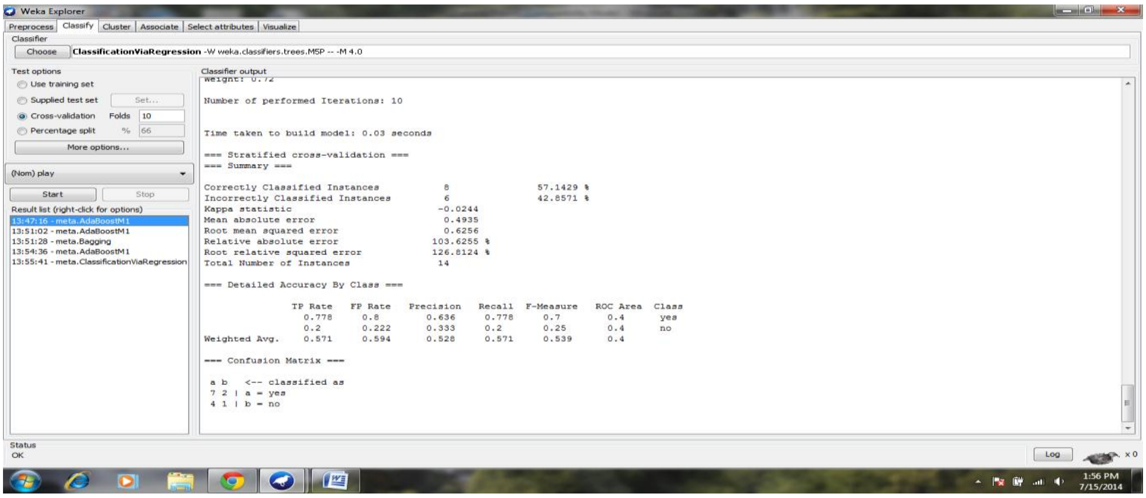
Status
OK

x 0

1:55 PM
7/15/2014

Figure 1: Screen shot for C4.5 classifier performance

2. Adaboost It is a machine algorithm, formulated by Yeave Freud and Robert Scapire. It is a meta-learning algorithm and used in conjunction with many other learning algorithms to improve their performance. [11]AdaBoost is an algorithm for constructing a"strong" classifier as linear combination. AdaBoost is adaptive only for this reason that subsequent classifiers built are weaked 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.



Classifier: ClassificationViaRegression -W weka.classifiers.trees.M5P -- -M 4.0

Test options: ○ Use training set ○ Supplied test set ● Cross-validation Folds 10 ○ Percentage split % 66

(Nom) play

Result list:
- 13:47:16 - meta.AdaBoostM1
- 13:51:02 - meta.AdaBoostM1
- 13:51:28 - meta.Bagging
- 13:54:36 - meta.AdaBoostM1
- 13:55:41 - meta.ClassificationViaRegression

```
Weight: 0.72

Number of performed Iterations: 10


Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances           8               57.1429 %
Incorrectly Classified Instances         6               42.8571 %
Kappa statistic                         -0.0244
Mean absolute error                      0.4935
Root mean squared error                  0.6256
Relative absolute error                103.6255 %
Root relative squared error            126.8124 %
Total Number of Instances               14

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.778     0.8        0.636     0.778     0.7        0.4      yes
                 0.2       0.222      0.333     0.2       0.25       0.4      no
Weighted Avg.    0.571     0.594      0.528     0.571     0.539      0.4

=== Confusion Matrix ===

 a b   <-- classified as
 7 2 | a = yes
 4 1 | b = no
```

Status: OK

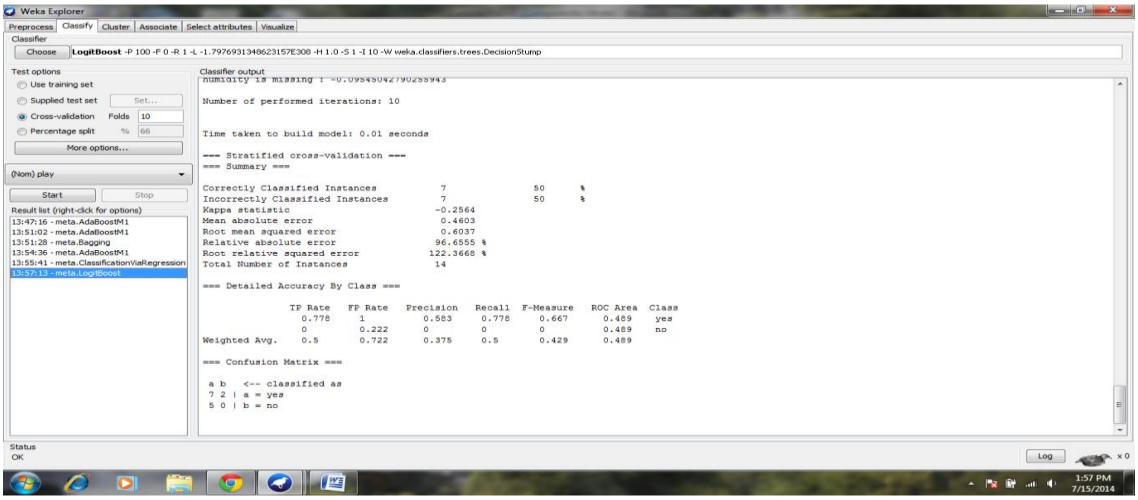
Figure 2: Screen shot for Adaboost classifier performance

3. Logiboost LogitBoost[11] is a boosting algorithm formulated by Jerome Friedmome, Trevor Hastie, and Robert Tibshirani. LogitBoost represents an application of established logistic regression to the AdaBoost method. Rather than minimizing error with respect to y, weak learners are chosen to minimize the (weighted least-squares) error of $f_t(x)$ with respect to

$$z_t = \frac{y^* - p_t(x)}{2p_t(x)(1 - p_t(x))},$$

where $p_t(x) = \frac{e^{F_{t-1}(x)}}{e^{F_{t-1}(x)} + e^{-F_{t-1}(x)}}$, $w_t = p_t(x)(1 - p_t(x))$ and $y^* = \frac{y+1}{2}$.

That is z_t is the Newton-Raphson approximation of the minimizer of the log-likelihood error at stage t, and the weak learner f_t is chosen as the learner that best approximates $z_t$by weighted least squares.



Classifier: LogitBoost -P 100 -F 0 -R 1 -L -1.7976931348623157E308 -H 1.0 -S 1 -I 10 -W weka.classifiers.trees.DecisionStump

Test options: ○ Use training set ○ Supplied test set ● Cross-validation Folds 10 ○ Percentage split % 66

(Nom) play

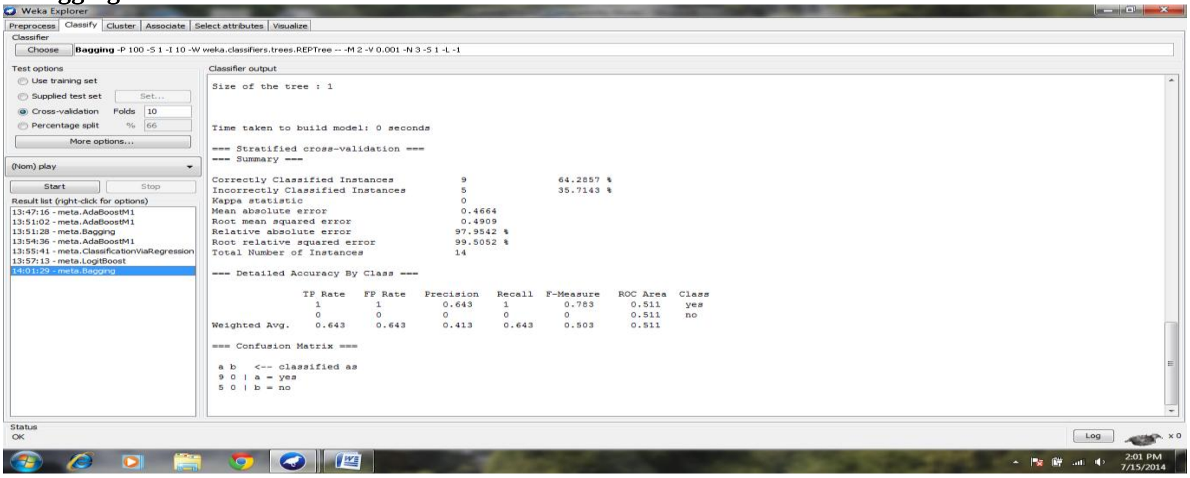
Result list:
- 13:47:16 - meta.AdaBoostM1
- 13:51:02 - meta.AdaBoostM1
- 13:51:28 - meta.Bagging
- 13:54:36 - meta.AdaBoostM1
- 13:55:41 - meta.ClassificationViaRegression
- 13:57:13 - meta.LogitBoost

```
humidity is missing : -0.09545042790255943

Number of performed iterations: 10


Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances           7               50      %
Incorrectly Classified Instances         7               50      %
Kappa statistic                         -0.2564
Mean absolute error                      0.4603
Root mean squared error                  0.6037
Relative absolute error                 96.6555 %
Root relative squared error            122.3668 %
Total Number of Instances               14

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.778     1          0.583     0.778     0.667      0.489    yes
                 0         0.222      0         0         0          0.489    no
Weighted Avg.    0.5       0.722      0.375     0.5       0.429      0.489

=== Confusion Matrix ===

 a b   <-- classified as
 7 2 | a = yes
 5 0 | b = no
```

Status: OK

Figure 3: Screen shot for Logiboost classifier performance

4. Bagging



Classifier: **Bagging** -P 100 -S 1 -I 10 -W weka.classifiers.trees.REPTree -- -M 2 -V 0.001 -N 3 -S 1 -L -1

Test options
○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) play

Result list (right-click for options)
13:47:16 - meta.AdaBoostM1
13:51:02 - meta.AdaBoostM1
13:51:28 - meta.Bagging
13:54:36 - meta.AdaBoostM1
13:55:41 - meta.ClassificationViaRegression
13:57:13 - meta.LogitBoost
14:01:29 - meta.Bagging

Classifier output

```
Size of the tree : 1



Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances           9               64.2857 %
Incorrectly Classified Instances         5               35.7143 %
Kappa statistic                          0
Mean absolute error                      0.4664
Root mean squared error                  0.4909
Relative absolute error                 97.9542 %
Root relative squared error             99.5052 %
Total Number of Instances               14

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         1         0.643     1         0.783      0.511    yes
                 0         0         0         0         0          0.511    no
Weighted Avg.    0.643     0.643     0.413     0.643     0.503      0.511

=== Confusion Matrix ===

 a b   <-- classified as
 9 0 | a = yes
 5 0 | b = no
```

Status
OK

2:01 PM
7/15/2014

Figure 4: Screen shot for Grading classifier performance

Algorithm classification	Correct Classification Rate	Mis- Classification
CART	88.0000	12.0000
ADABOOST	73.0000	27.0000
GRADING	54.3333	45.6667
BAGGING	90.3333	9.6667

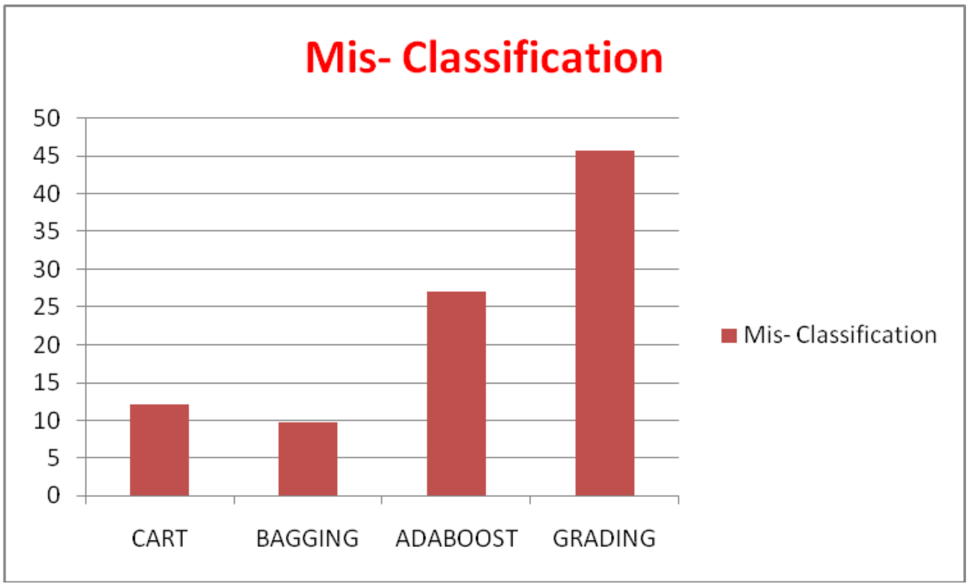


Figure-4

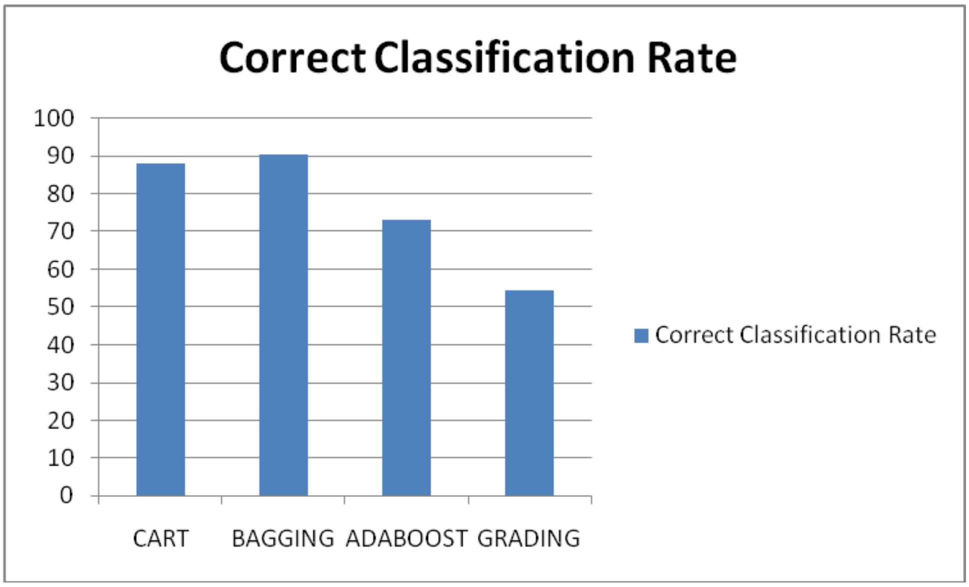


Figure-5

.Graph for Correct Classification VS. Misclassification rate

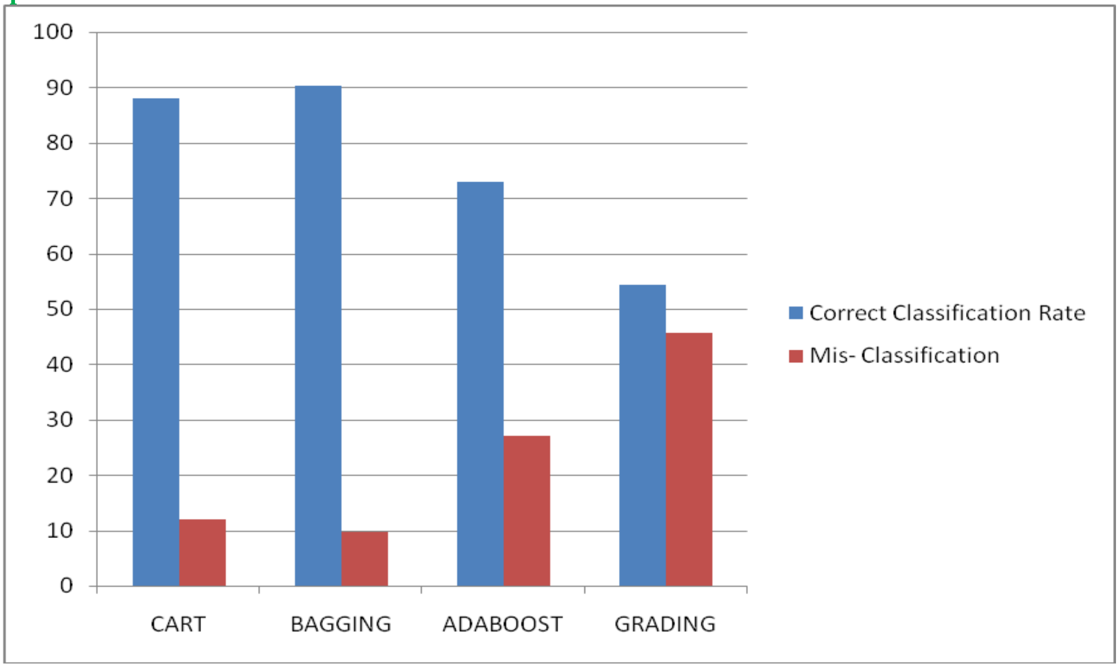


Figure-6

III. CONCLUSIONS

This paper represents computational issues of four supervised machine learning algorithms i.e., Classification and Regression technique algorithm, Adaboost algorithm, Logitboost algorithm and Bagging algorithm with dedicating role of detection of weather forecast on the basis of classification rule. Among four algorithms, Bagging algorithm is the best in performance.. In order to compare the classification performance of four machine learning algorithms, 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 concluded that Baging is the best as compared to adaboost, logitboost, Classification and Regression Technique.

REFERENCES

[1] QUINLAN, J. R. (1988). C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, San Mateo, CA.

[2] C4.5 Algorithm - Disponible inhttp://www.cse.unsw.edu.au/~quinlan . Accessed in 25 July 2010.

[3] Bruno Carneiro da Rocha1,2 and Rafael Timóteo de Sousa Júnior2 "IDENTIFYING BANK FRAUDS USING CRISP-DM AND DECISION TREES" International journal of computer science & information Technology (IJCSIT) Vol.2, No.5, October 2010.

[4] Mabroukeh, N.R. and C.I. Ezeife, 2010. A taxonomy of sequential pattern mining algorithms. ACM Comput. Surveys. DOI: 10.1145/1824795.1824798

[5] Ramageri, B.M. and B.L. Desai, 2013 Role of datamining in retail sector. Int. J. Comput. Sci. Eng., 5.
Moradi, M., M. Salehi, M.E. Ghorgani and H.S. Yazdi, 2013. Financial distress prediction of Iranian companies by using data mining techniques. Organizacija, 46: 20-27.

[6] Ramageri, B.M. and B.L. Desai, 2013 Role of data mining in retail sector. Int. J. Comput. Sci. Eng., 5.

[7] Delamaire, L., A. Hussein and P. John, 2009. Credit card fraud and detection techniques: A review. Banks Bank Syst., 4: 57-68.

[8] Ravisankar, P., V. Ravi, G.R. Rao and I. Bose, 2011. Detection of financial statement fraud and feature selection using data mining techniques. Decision Support Syst., 50: 491-500. DOI: 10.1016/j.dss.2010.11.006.

[9] Raj, S.B.E. and A.A. Portia, 2011. Analysis on credit card fraud detection methods. Proceedings of the International Conference on Computer Communication and Electrical Technology, Mar. 18- 19, IEEE Xplore Press, Tamilnadu, pp: 152-156.DOI: 10.1109/ICCCET.2011.5762457.

[10] Petry, F.E. and L. Zhao, 2009. Data mining by attribute generalization with fuzzy hierarchies in fuzzy databases. Fuzzy Sets Syst., 160: 2206-2223. DOI 10.1016/j.fss.2009.02.014.

[11] Changdola, V., A. Banerjee and V. Kumar, 2009. Anomaly detection: A survey. ACM Comput. Surveys, 9: 1-72.

[12] Moin, K.I. and Q.B. Ahmed, 2012. Use of data mining in banking. Int. J. Eng. Res. Applic., 2: 738-742.

[13] Chen, I., L. Chi-Jie, L. Tian-Shyug and L. Chung-Ta, 2009. Behavioral scoring model for bank customers using data envelopment analysis. Stud. Comput. Intell., 214: 99-104. DOI: 10.1007/978-3-540-92814-0_16
Yap, B.W., S.H. Ong and N.H.M. Hussain, 2011. Using

[14] Yap, B.W., S.H. Ong and N.H.M. Hussain, 2011. Using data mining to improve assessment of credit worthiness via credit scoring models. Expert Syst. Appli., 38: 13274-13283. DOI: 10.1016/j.eswa.2011.04.147.

[15] Bhattacharya, S., S. Jha, K. Tharakunnel and J.C. Westland, 2011. Data mining for credit card fraud: A comparative study. Decision Support Syst., 50: 602- 613. DOI: 10.1016/j.dss.2010.08.008 .

[16] Ravisankar, P., V. Ravi, G.R. Rao and I. Bose, 2011. Detection of financial statement fraud and feature selection using data mining techniques. Decision Support Syst., 50: 491-500. DOI:

[17] Ravisankar, P., V. Ravi, G.R. Rao and I. Bose, 2011.Detection of financial statement fraud and feature selection using data mining techniques. Decision Support Syst., 50: 491-500. DOI: 10.1016/j.dss.2010.11.006 .

[18] Dorr, D.M. and M.D. Anne, 2009. Establishing relationships among patterns instock market data. Data Knowl. Eng., 68: 318-337. DOI 10.1016/j.datak.2008.10.001.

[19] Tsai, H.H., 2012. Global data mining: An empirical study of current trends, future forecasts and technology diffusions. Expert Syst. Applic., 39: 8172-8181.