ISSN: 2321-7782 (Online)
Volume 2, Issue 12, December 2014
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

Application of Meta Learning Algorithms for the Prediction of Diabetes Disease

Sanjay Kumar Sen[1]
Assistant Professor, Dept. of CSE
Orissa Engineering College
Bhubaneswar, Odisha – India

Dr. Sujata Dash[2]
Reader, PG Dept. of Computer Application
North Orissa University
Bhubaneswar, Odisha – India

Abstract: Preventing the disease of diabetes becomes a challenging factor to the healthcare community. So medical professionals need a reliable prediction methodology to diagnose Diabetes. Although many studies employ several data mining techniques to assess the leading causes of diabetes, only small sets of clinical risk factors are considered. It is very much need for a medical professionals for a reliable production methodology to diagnose diabetes. The proposed work in this paper is the combination of four supervised machine learning algorithms, Classification and Regression Tree (CART), Adaboost algorithm, Logiboost algorithm, Grading algorithm. The experimental result shows the performance analysis of different meta-learning algorithms and also compared on the basis of misclassification and correct classification rate, the error rate focuses True Positive, True Negative, False Positive and False Negative and Accuracy. This project aims for mining the relationship in Diabetes data for efficient classification. The data mining methods and techniques will be explored to identify the suitable methods and techniques for efficient classification of Diabetes dataset and in mining useful patterns.

Key words- machine learning algorithm, data mining.

I. INTRODUCTION

Now days Diabetes disease is rising rapidly in the world and becomes one of the most common diseases in the world which becomes a major health problem in the world. Diabetes is often called a modern-society disease. Diabetes, also called as diabetes mellitus, is a process of a group of metabolic diseases in which the person has high blood glucose also called blood sugar, either due to inadequate production of insulin, or due to the body's cells do not respond properly to insulin, or both. Patients with high blood sugar will suffer from polyuria (frequent urination), for which they will become increasingly thirsty (polydipsia) and hungry (polyphagia). Basically **t**here are three types of diabetes: 1) Type 1 Diabetes in which the body does not produce adequate insulin..This type of type 1 diabetes develop below the age of 40, often in early adulthood or teenage years.. Approximately 10% of all diabetes cases are type 1. Patients having type 1 diabetes will need to take insulin injections for the rest of their life. They shuld also maintain proper blood-glucose levels by carrying out time to time regular blood tests and with a special diet and regular exercise.2) Type 2 Diabetes in which either the body does not produce sufficient insulin for proper function, or the cells in the body do not absorb the insulin. About 90% of all cases of diabetes worldwide are of this type. This type disease may be controlled by loosing weight, taking a healthy diet, doing plenty of exercise, and monitoring their blood glucose levels. However, type 2 diabetes is typically a progressive disease - it gradually gets worse - and the patient will probably end up have to take insulin, usually in tablet form. Overweight, obese people, belly fat have a chance of developing type 2 diabetes compared to those with a healthy body weight.. **3) Gestational Diabetes** in this type it affects females during pregnancy. Women having very high levels of glucose in their blood, as a result bodies are unable to produce enough insulin to transport all of the glucose into their cells, resulting in progressively rising levels of glucose. There is a risk of complications during childbirth. There is an abnormal size of baby. It can be controlled by adequate by adequate diet and exercise.. According

© 2014, IJARCSMS All Rights Reserved

to data from the 2011 National Diabetes Fact Sheet[1] 25.8 million people, or 8.3% of the U.S. population, have diabetes. The estimated total cost of diabetes in the United States for 2007 was $174 billion. Worldwide, the picture is very similar, with an estimated 285 million people affected by diabetes in 2010, representing 6.6% of the world's adult population. Health care expenditures for diabetes are expected to be $490 billion for 2030, accounting for 11.6% of the total health care expenditure in the world.[2]Diabetes is a disease in which the human body is unable to produce the adequate amount of insulin needed to regulate the amount of sugar. Classification Algorithms usually require that the classes be defined based on the data attribute values. They often describe these classes by looking at the characteristics of data already known to belong to class. Pattern Recognition is a type of classification where an input pattern is classified into one of the several classes based on its similarity to these predefined classes. Knowledge discovery in databases denotes the complex process of identifying valid, novel, potentially useful and ultimately understandable patterns in data [4]. A common methodology for distributed machine learning and data mining is of two-stage, first performing local data analysis and then combining the local results forming the global one [5]. For example, in [6], a meta-learning process was proposed as an additional learning process for combining a set of locally learned classifiers (decision trees in particular) for a global classifier. Knowledge Discovery in Databases (KDD) is the process of finding useful information and patterns in data which involves Selection, Pre-processing, Transformation, Data Mining and Evaluation[3]. Meta-learning[12] is loosely defined as learning from learned knowledge. It is a recent technique that seeks to compute higher level models, called meta-classifiers, that integrate in some principled fashion the information cleaned by the separately learned classifiers to improve predictive performance. In meta learning process a number of learning programme is executed on a number of data subsets in parallel then collective result is collected in the form of classifiers

II. PROPOSED SYSTEM

We have applied meta learning algorithm to classify Diabetes Clinical data and predicts whether a patient is affected with Diabetes or not. The dataset used for this purpose is of Pima Indians Diabetes Database of National Institute of Diabetes and Digestive and Kidney Diseases from UCI Machine Learning Repository [7]. The dataset contains 9 attributes having 768 instance samples, This dataset is for classification purpose applying various meta learning algorithms.

III. ALGORITHM USED

a. 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].

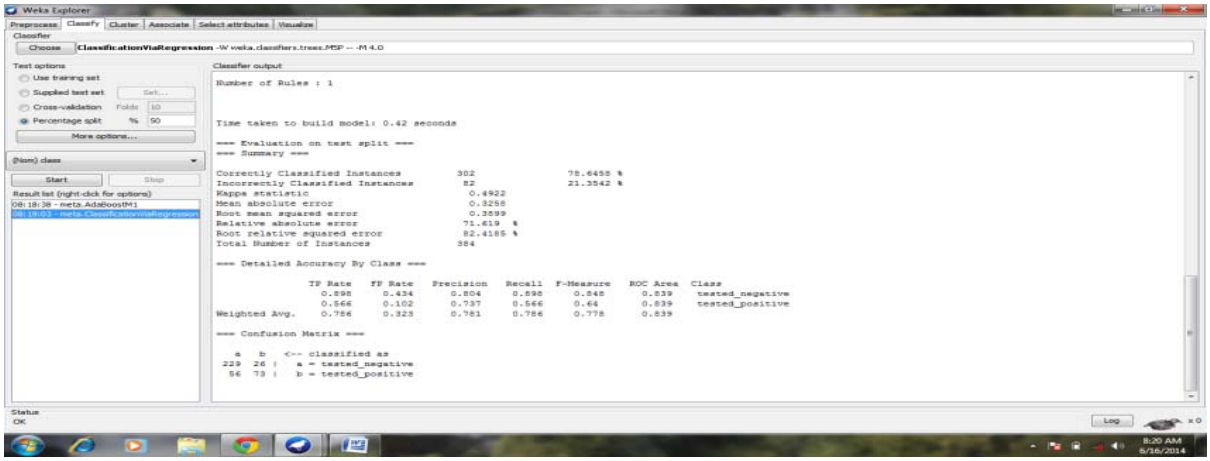

Figure 1: Screen shot for C4.5 classifier performance

© 2014, IJARCSMS All Rights Reserved

b. 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

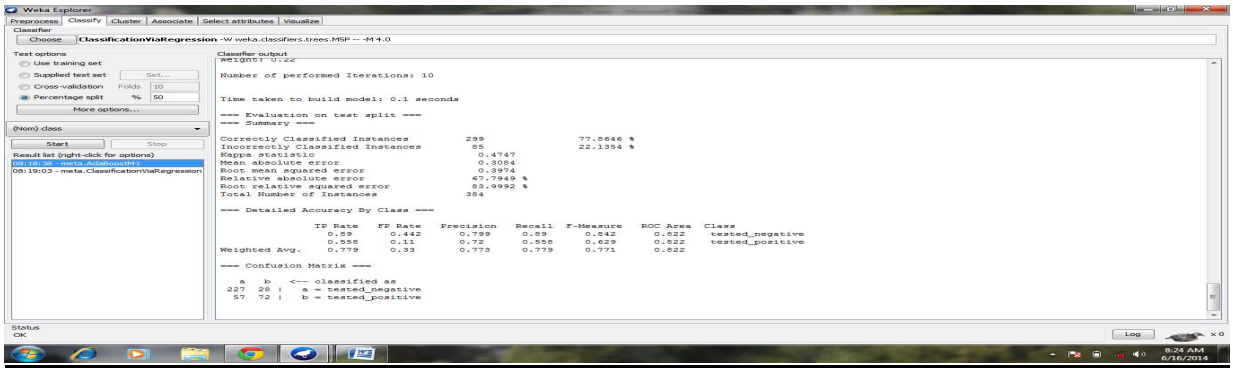

Figure 2: Screen shot for Adaboost classifier performance

c. 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.

Figure 3: Screen shot for Logiboost classifier performance

d. Grading

© 2014, IJARCSMS All Rights Reserved

Grading algorithm is used to learn for prediction of each of the original learning algorithms whether its prediction for a particular example is correct or not. We 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)[10] who propose to use a meta-classification scheme for characterizing model errors.. The algorithm is also used as an attempt to extend the work of Bay and Pazzani (2000)[9] who propose to use a meta-classification scheme for characterizing model errors. [11]Their suggestion is to learn a comprehensible theory that describes the regions of errors of a given classifier.

Figure 4: Screen shot for Grading classifier performance

IV. PERFORMANCE COMPARISONS

Algorithm classification	Correct Classification Rate	Mis- Classification Rate
CART	78.646	21.354
ADABOOST	77.864	22.136
LOGIBOOST	77.479	22.521
GRADING	66.406	33.594

Table -1

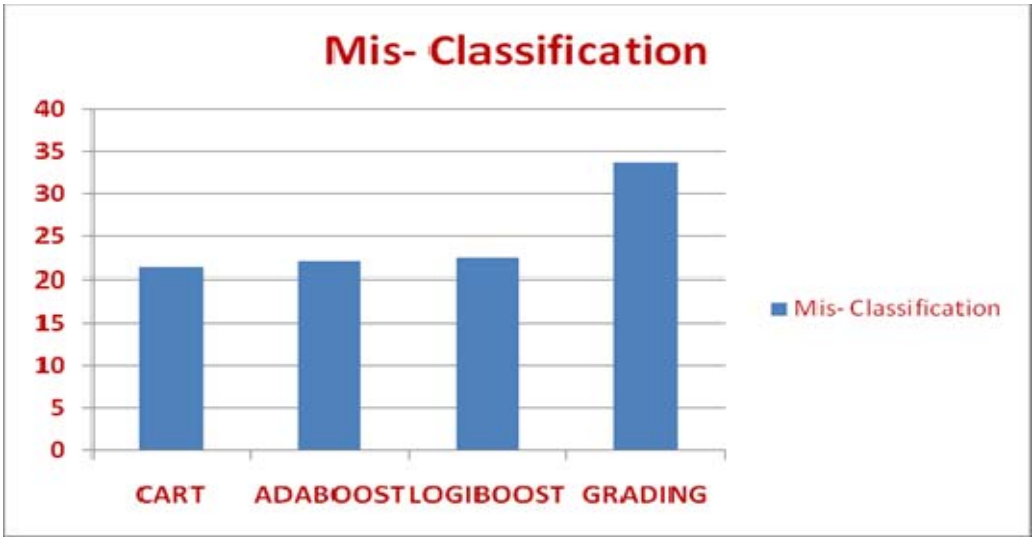

Figure-4

© 2014, IJARCSMS All Rights Reserved

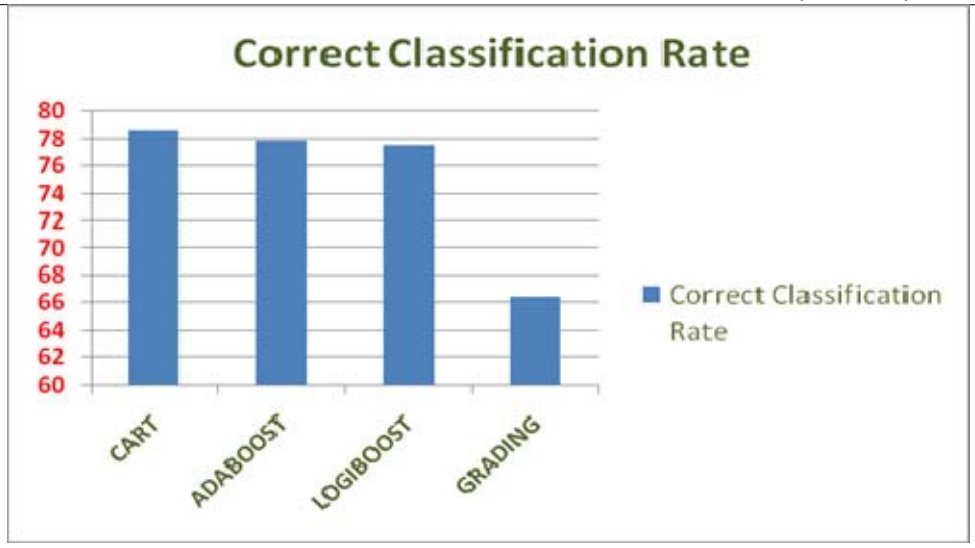

Figure-5

Graph for Correct Classification VS. Misclassification rate

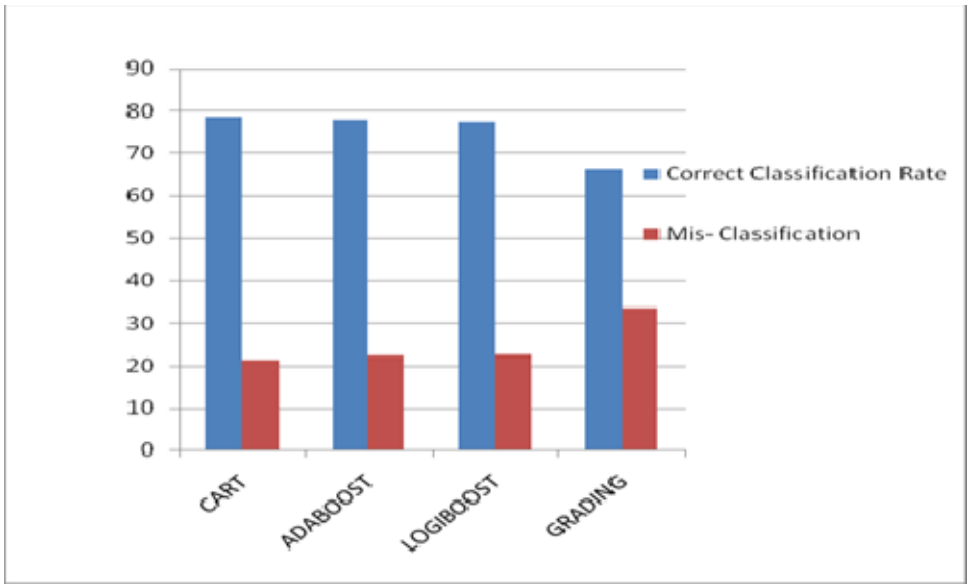

Figure-6

V. CONCLUSION

This paper represents computational issues of four supervised machine learning algorithms i.e., Classification and Regression technique algorithm, Adaboost algorithm, Logitboost algorithm and Grading algorithm with dedicating role of detection of Diabetes disease on the basis of classification rule. Among four algorithms, Classification and Regression technique algorithm is the best because the Classification and Regression technique algorithm is easier to interpret and understand as compared to Adaboost, Logitboost and Grading algorithm. 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 Classification and Regression technique algorithm is the best as compared to adaboost, logitboost, Grading algorithm.

References

1. Centers for Disease Control and Prevention. National diabetes fact sheet: national estimates and general information on diabetes and prediabetes in the United States, 2011. Available from:http://www.cdc.gov/diabetes/pubs/pdf/ndfs_2011.pdf. Accessed April, 1 2011.
2. IDF Diabetes Atlas. The Economic Impacts of Diabetes. Available from:http://www.diabetesatlas.org/content/economic-impacts-diabetes. Accessed April 1, 2011.
3. K. Rajesh, V. Sangeetha "Application of Data Mining Methods and Techniques for Diabetes Diagnosis" International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 3, September 2012.

© 2014, IJARCSMS All Rights Reserved

4. R. Grossman, S. Baily, S. Kasif, D. Mon, and A. Ramu. The preliminary design of papyrus: A system for high performance. In P. Chan H. Kargupta, editor, Work. Notes KDD-98 Workshop on Distributed Data Mining, pages 37–43. AAAI Press, 1998.

5. Zhang X., Lam C., Cheung W.K., Mining Local Data Sources For Learning Global Cluster Model Via Local Model Exchange. IEEE ntelligence Informatics Bulletin, (4) 2 (2004).

6. Prodromidis A., Chan P.K., Stolfo S.J., Meta-learning in Distributed Data Mining Systems:Issues and Approaches. In: Advances in Distributed and Parallel Knowledge Discovery, Kargupta H., Chan P.(ed.), AAAI/MIT Press, Chapter 3, (2000).

7. UCI Machine Learning Repository- Center for Machine Learning and Intelligent System, http://archive.ics.uci.edu..

8. P. Chan and S. Stolfo. Meta-learning formulti strategy and parallel learning. InProcSecondIntl. Work. Multistrategy Learning, pages 150–165, 1993.

9. 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.

10. 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.

11. Sanjay Kumar Sen, Prof. Dr Sujata Dash " Meta Learning Algorithms for Credit Card Fraud Detection" IJERD ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com Volume 6, Issue 6 (March 2013), PP. 16-20

12. 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.

13. 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.

© 2014, IJARCSMS All Rights Reserved