Volume 6, Issue 2, February 2016 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

An Experimental Analysis of Different Classification Techniques for Diabetes Dataset

Kamalakkannan V
M Phil Research Scholar
Computer Science Department
Bharathiar University, Coimbatore,
Tamil Nadu, India

Dr. Ramyachitra D
Assitant Professor
Computer Science Department
Bharathiar University, Coimbatore,
Tamil Nadu, India

Abstract— The classification is one of the fundamental cognitive processes used to classify and apply our knowledge. The present study designed to do the appearance analysis of several classification techniques using machine learning tools. By using the classification techniques have been tested a diabetes datasets. In this article the classifier algorithms namely BayesNet, SMO, Bagging algorithms are analyzed to predict which is the best classification algorithm for Diabetes Diagnosis datasets. The experiments are prepared using the 10 fold Cross-validation method. For the comparative analysis the performance metrics such as execution time, classification accuracy and error rates are used for analyzing the results. From this study it is found that the best performance algorithm for the Diabetes dataset.

Keywords— BayesNet, Sequential Minimal Optimization (SMO), Bagging, Classification, WEKA, diabetes dataset.

I. INTRODUCTION

The classification task consists in assigning instances from a given domain, described by a set of discrete valued attributes, into a set of classes, which can be considered values of a selected discrete target attribute called target concept. The correct class labels are generally unknown, but are provided for a subset of the domain. It can be used to create the classification model, which is a machine friendly representation of the knowledge needed to classify any possible instance from the same domain, describes by the same set of attributes. This follows the general assumption of inductive learning of the classification task is the most common instantiation [1].

The classification Techniques disperses a class to collection of data records having specific attributes and its values. Thus the classification techniques in healthcare can be applied for diagnostics purposes. A classification model receives a set of related attribute values, such as clinical measurements and gives a class of data records as output [2].

In this paper comparison has been done with machine learning algorithms namely Bayesian Network, Sequential Minimal optimization (SMO) and Bagging to find out which classifier algorithms is suitable for the diabetes dataset. In the testing option there are four types of parameters such as training set, percentage split, cross validation, and supplied test set. Section 2 describes the literature review, Section 3 describes the methodology for the Diabetes Diagnosis dataset and section 4 discusses our experimental analysis for Diabetes dataset. And finally Section 5 gives the conclusion.

II. LITERATURE REVIEW

Mukesh kumari, et al., focused the Bayesian Network classification algorithm for Diabetes using Weka tool. They had taken 206 instances and 9 attributes Diabetes as input. They used the Diabetes dataset for their experimental results based on accuracy and error rate performances. They conclude that Bayesian Network is a best classification technique based on performance metrics [3].

Rashedur M. Rahman et al., discussed different classification algorithms such as MLP, BayesNet, J48graft, JRip, FLP. The performances of the classifiers were measured and then the results are compared with accuracy and error rate using the Diabetes Diagnosis dataset. They conclude that Bayesian Network classification algorithm provides better accuracy and error rate when compare with other classification algorithms [4].

R. Sujatha, determined the different classification techniques to produces the excellent prediction for Diabetes, Soy Beans and Wheat seed datasets. They used the various classification techniques are ZeroR, OneR, Decision Table, Naïve Bayes, PART, SMO, J48, Random Tree. Then the results are evaluated based on the accuracy, paired T test and statistical methods. They decided SMO is the best classification algorithm for those datasets when compared with other algorithms in their research [5].

Gouda I. Salama et al., performs a comparative analysis among the various classifiers such as J48, MLP, Naïve Bayes, SMO, and IBK by using classification accuracy and confusion matrix based on 10 fold Cross-validation for Breast Cancer Dataset. They determined the fusion of SMO and IBK is better to the other classifiers algorithms [6].

Najmeh Hosseinpour et al., discussed the efficiency of classification algorithms used are Bagging, AdaBoost, Random Forest and Multiclass Classifier. They used multiple classification techniques for diabetes diagnosis datasets. The analyses of experimental results are compared with accuracy. Thus the evaluation of results indicates that bagging with logistic core has best performs well for diabetes diagnosis datasets [7].

M. C. Tu et al., focused a simple classification technique and they used Cleveland Heart Disease dataset. The performances of the classifiers were measured and then results are compared with accuracy, sensitivity and specificity using 10 fold Cross-validation to test the data. They conclude that Bagging is the best classification algorithm for these datasets in their research [8].

III. METHODOLOGY

The different classification techniques are used to find the best algorithm for Diabetes dataset based on the 10 fold Cross-validation. The comparative analysis diagram is given below Fig. 1:

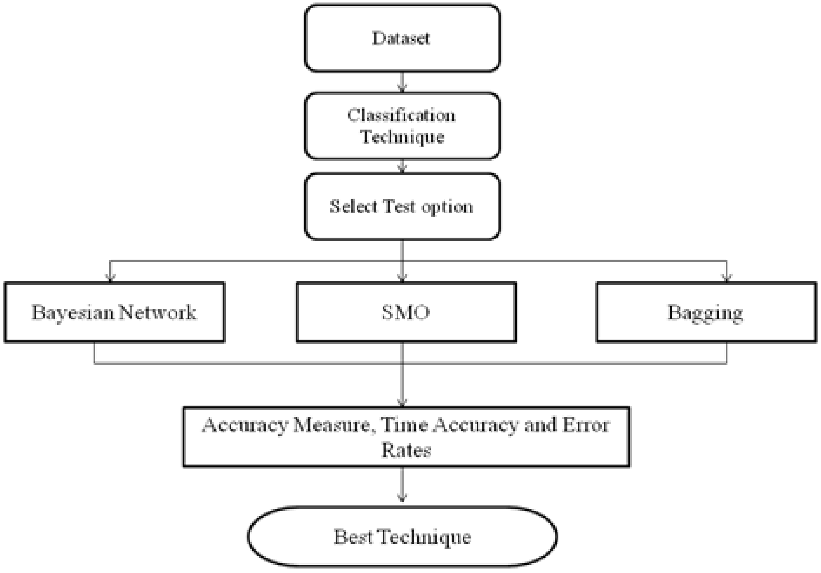

Fig. 1: Stream Plan for Virtual Analysis.

The classification is one of the techniques in Data Mining solves various problems like algorithm selection, division of training, testing data and model comparison. It is two stairs, first build classification model using training data. Every object of the data must be pre-classified. After, the model generated in the preceding step is tested by assigning class labels to data objects in a test dataset. Thus the test data is different from the training data. Thus the accuracy of the classification model is determined by comparing true class labels in the testing set with those assigned by the model [9].

In this paper we using 3 different classification Techniques to finding the better algorithm for Diabetes Diagnosis dataset and the methods are as follows,

A. Bayes Net

The BayesNet learns Bayesian networks under the assumptions made nominal attributes and no missing values. A Bayesian network is a structure that shows the conditional dependencies between domain variable and may also be used to illustrate graphically the probabilistic causal relationships among domain variables. A Bayesian network consists of a directed acyclic graph and probability tables. The nodes of the network represent the domain variables and an arc between two nodes indicates the existence of a causal relationship among these two nodes. Associated with each node there exist a probability table. Although the domain variables can be continuous, they are discretized most of the time for simplicity and efficiency. Besides representing the dependencies between domain variables, a Bayesian network is used for inferencing the probability of a variable given the observation of other variables. The implementation of the Bayesian networks in Weka [10].

B. The Sequential Minimal Optimization

The SMO class implements the Sequential Minimal Optimization, which learns this kind of classifier. The SMO is a new algorithm for training Support Vector Machines (SVM). The fastest methods for learning support vector machinery, sequential minimal optimization is often slow to converge to a solution particularly when the data is not linearly separable in the space spanned by the nonlinear mapping. This frequently happens, because of noise data. Both run time and accuracy depend critically on the values that are given to two parameters: such as the upper bound on the coefficients values in the equation for the hyperplane, and next one is degree of the values in the non-linear mapping then both are default values set to one. The best setting for a particular dataset can be found only by experiment [11].

C. The Bagging

The Bagging is method that combines bootstrapping and aggregating. It is general and ensemble technique for data classification and forecast. An enhanced complex classifier is created in order to improve the classification accuracy as compared to produce a final output class. All individual classifiers are qualified from 10 folds Cross-validation by discussion sample of instances with substitution. The size of each sample and 10 folds Cross-validation is equal [12].

IV. EXPERIMENTAL RESULTS

In this experimental analysis the Diabetes Diagnosis dataset is taken as input for classification. This research used the Weka open source data mining tools for analyzing the Diabetes Diagnosis data.

A. The Dataset

The Diabetes Diagnosis dataset has collected from UCI ML (Machine Learning) Repository [13]. The Diabetes Diagnosis dataset contains 768 instances and 9 attributes. The evaluation has been done on cross - validation parameter to test the performance of the three classification algorithms to determine the best technique for the Diabetes Diagnosis dataset by using the performance factors such as accuracy, time estimation and error rate.

B. The Accuracy and Statistical Analysis

By using the 10 fold cross-validation option to predict the correctly classified and incorrectly classified instances for the classifiers shown in Table 1. The BayesNet, SMO, Bagging classifiers are compared with accuracy measures that are depicted in fig 2. Therefore, the SMO classification algorithm has high accuracy to compare with BayesNet and Bagging algorithms for Diabetes Diagnosis dataset.

Table I Classified Instances for Diabetes Dataset

Algorithms	BayesNet	SMO	Bagging
Correctly Classified	74.35	77.34	75.26
Incorrectly Classified	25.65	22.66	24.74

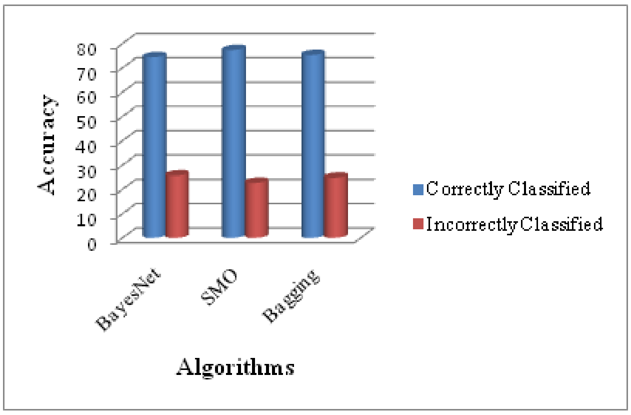

Fig. 2 Accuracy Measure For Classification Techniques

For correctly classified instances, it is inferred that SMO algorithms performs 3.87% better than the BayesNet algorithm and 2.69% better than the Bagging algorithm. Similarly for incorrectly classified instances it is inferred that SMO algorithm performs 11.66% better than the BayesNet algorithm and 22.09% better than Bagging algorithm.

Table 2 shows the performance of the time estimation for BayesNet, SMO and Bagging classification algorithms. It is inferred that the SMO has least time taken when compared with BayesNet and Bagging algorithms.

Table III Time Execution For Classification Techniques

Algorithms	Execution Time
BayesNet	0.05
SMO	0.03
Bagging	0.31

Thus the TP Rate, FP Rate, F-Measure, ROC area and Kappa Statistics values had been shown in the table 3. It is inferred that for the SMO classification algorithm the TP Rate, FP Rate, F-Measure, and Kappa Statistics Values are higher when compared with BayesNet and Bagging classification algorithms using the 10 fold cross-validation parameter. But the ROC area is less than the BayesNet and Bagging classification algorithms for Diabetes Diagnosis dataset. From these experimental results the SMO classification algorithm performs better when compared with BayesNet and Bagging classification algorithms that are depicted in Fig 3.

Table IIIII Performance Metrics of classification Techniques.

Algorithms	TP Rate	FP Rate	F-Measure	ROC Area	Kappa Statistics
BayesNet	0.74	0.32	0.74	0.8	0.43
SMO	0.77	0.33	0.76	0.72	0.47
Bagging	0.75	0.32	0.75	0.75	0.44

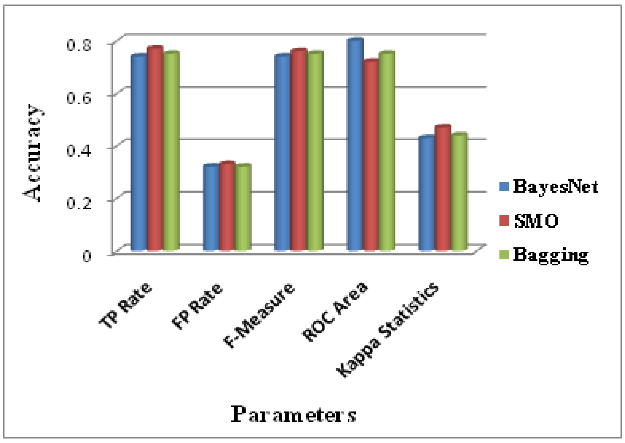

Fig. 3 Comparison of Performance Metrics For Diabetes Dataset.

The error rate of the classification algorithms are depicted in the table 4. The error rates are as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Relative Absolute Error (RAE) and Root Relative Squared Error (RRSE). These error rate measures are compared for Diabetes Diagnosis dataset. From that it is observed that the SMO had high error rate for RMSE, RRSE and least error rate for MAE, RAE as compared with other algorithms in the fig 4&5.

Table IVV Error Rate Measures for Classification Techniques

Algorithms	**MAE**	**RMSE**	**RAE**	**RRSE**
BayesNet	0.2987	0.42	65.71	88.28
SMO	0.2266	0.476	49.85	99.86
Bagging	0.3161	0.4076	69.56	85.52

For MAE, it is inferred that SMO algorithm is 24.14% better than the BayesNet algorithm and 28.31% better than the Bagging algorithm. For RMSE it is inferred that SMO algorithm performs 11.76% better than the BayesNet algorithm and 14.37% better than the Bagging algorithm. For RAE it is inferred that SMO algorithm performs 24.14% better than the BayesNet algorithm and 28.34% better than the Bagging algorithm. For RRSE it is inferred that SMO algorithm is 11.6% better than the BayesNet algorithm and 14.36% better than the Bagging algorithm.

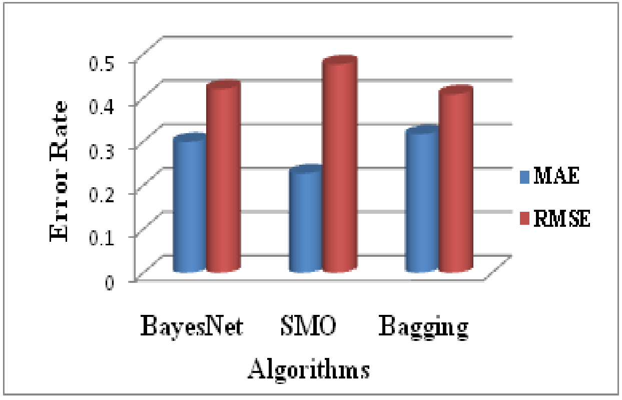

Fig.4 Error Rate Of MAE & RMSE For Diabetes Dataset

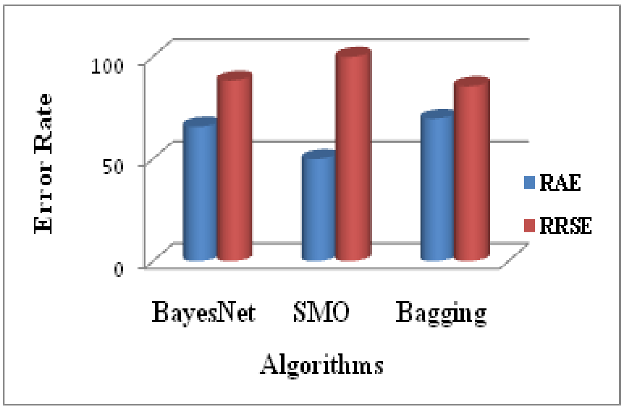

Fig. 5 Error Rate Of RAE & RRSE For Diabetes Dataset.

V. CONCLUSION

In this paper different classification algorithms such as the Bayesian Network, Sequential Minimal Optimizations (SMO) and Bagging are experimentally evaluated using Diabetes Diagnosis dataset. The classification algorithms analyses are based on accuracy, execution time and error rates using the data mining tool as Weka. The 10 fold cross-validation testing parameter is used for the experimental analysis. Form the outcome it is inferred that SMO algorithm provides better results for Diabetes Diagnosis dataset when compared with other classification algorithms. In future the performance of SMO classifier can be implemented on other datasets also. In future the SMO algorithm can be hybridized to obtain more effective results.

REFERENCE

[1] Ian H. Witten, Eibe Frank, Mark A. Hall,"Data Mining Practical Machine Learning Tools and Techniques",3rd Edition, Morgan Kaufmann Publishers is an imprint of Elsevier 30 Corporate Drive, Suite 400, Burlington, MA 01803, USA

[2] Vidyullatha Pellakuri1, Deepthi Gurram, Dr.D. Rajeswara Rao, Dr.M.R.Narasinga Rao, " Performance Analysis and Optimization of Supervised Learning Techniques for Medical Diagnosis Using Open Source Tools", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (1) , 2015, 380-383.

[3] Mukesh kumari, Dr. Rajan Vohra, Anshul arora, "Prediction of Diabetes Using Bayesian Network", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (4) , 2014, 5174-5178.

[4] Rashedur M. Rahman, Farhana Afroz, "Comparison of Various Classification Techniques Using Different Data Mining Tools for Diabetes Diagnosis", Journal of Software Engineering and Applications, 2013, 6, 85-97.

[5] R.Sujatha, D.Ezhilmaran, "Evaluation of Classifiers to Enhance Model Selection", International Journal of Computer Science & Engineering Technology (IJCSET) ISSN: 2229-3345 Vol. 4 No. 01 Jan 2013.

[6] Gouda I. Salama, M.B.Abdelhalim, Magdy Abd-elghany Zeid, "Breast Cancer Diagnosis on Three Different Datasets Using Multi-Classifiers", International Journal of Computer and Information Technology (2277 – 0764) Volume 01– Issue 01, September 2012.

[7] Najmeh Hosseinpour, Saeed Setayeshi, Karim Ansari-asl, Mohammad Mosleh, "Diabetes Diagnosis by Using Computational Intelligence Algorithms", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 12, December 2012 ISSN: 2277 128X.

[8] M. C. Tu, D. Shin, and D. Shin, "Effective Diagnosis of Heart Disease through Bagging Approach," Biomedical Engineering and Informatics, IEEE, 2009.

[9] Nikita Bhatt, Amit Thakkar, Amit Ganatra, "A Survey & Current Research Challenges in Meta Learning Approaches based on Dataset Characteristics", International Journal of Soft Computing and Engineering (IJSCE), ISSN: 2231-2307, Volume-2, Issue-1, March 2012.

[10] Rashedur M. Rahman, Farhana Afroz, "Comparison of Various Classification Techniques Using Different Data Mining Tools for Diabetes Diagnosis", Journal of Software Engineering and Applications, 2013, 6, 85-97.

[11] Eibe Frank, Ian H. Witten,"WEKA Machine Learning Algorithms in Java", Morgan Kaufmann Publishers.

[12] Esteban Alfaro, Matias Gamez, Noelia Garcia, "adabag: An R Package for Classi_cation with Boosting and Bagging", Journal of Statistical Software, August 2013, Volume 54, Issue 2.

[13] C. Blake and C.J. Merz, UCI Repository of machine learning databases, University of California.