Volume 5, Issue 10, October-2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

An Efficient Bayes Classification Algorithm for analysis of Breast Cancer Dataset using Cross Validation Parameter

S. Kalaivani
Department of Computer Science
PGP Atrs and Science College
Namakkal, Tamil Nadu, India

S. Gandhimathi
Head of the Department
PGP Arts and Science College
Namakkal, Tamil Nadu, India

*Abstract –**Classification is one of the techniques used in the data mining field. Normally, the classification technique is used to predict group membership for data instances. Data mining also includes analysis and prediction for the data stored in the database. One of the data analysis task is a classification, where a model or classifier is used to construct or predict categorical labels. Classification technique is capable of processing a wider variety of data than regression and is growing in popularity. Classification techniques include several techniques such as decision trees, neural networks, etc. That is the models that accurately predict the class labels of previously unknown records. In this paper we are analyzing the performance of 3 classifiers algorithms namely Naïve Bayes, Random Tree and Support Vector Machine (SVM). For the comparison of different classification algorithms, we used the Breast Cancer dataset. And finally we find out the comparative analysis based on the performance factors such as the classification accuracy is performed on all the algorithms.***

Key words - Classification, Navie Bayes, Random Tree, Support Vector Machine (SVM), Breast Cancer Data set, Cross Validation.

I. INTRODUCTION

Generally Classification technique is one kind of predictive modeling. More specifically, classification is the process of assigning new objects to predefined categories or classes. The classification technique is used to predict group of membership for data instances [1].The accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified by a classifier. The associated class label of each test tuple is compared with the learned classifier's class prediction for that tuple. It describes several methods of estimating classifier accuracy. The classifier can be used to classify the future data tuples for which the class label is not known.

In this paper an analysis is made to find out which test option is the best for classifier algorithm named Naïve Bayes, Random Tree and Support Vector Machine (SVM). This paper uses the breast cancer dataset for comparison of those algorithms. And our paper is structured as follows. Section 2 describes the literature review, Section 3 describes the methodology for the breast cancer dataset and Section 4 describes our experimental result. And finally Section 5 gives the conclusion and future work.

II. LITERATURE REVIEW

Ron Kohavi , et al., scaled up the accuracy of Naive-Bayes Classifiers and a Decision-Tree Hybrid in retains the interpretability of Naive-Bayes and decision trees, while resulting in classifiers that frequently outperform both constituents, especially in the larger databases tested [2].

Jyoti Soni, et al., presented an overview of Heart Disease Prediction in a lack of effective analysis tools to discover hidden relationships and trends in data [3].

Jameela,et al., compared the decision and random tree algorithms to a web log data for finding frequent patterns. The preprocessed data is classified using classification algorithms like decision and random tree[4].

Virendra Raghuwanshi, et al., presented an Anomaly Base Network Intrusion Detection by Using Random Decision Tree and Random Projection. They used the NIDSs is to protect the resources from threats. It analyzes and predicts the behaviors of users, and then these behaviors will be considered an attack or a normal behavior use Random projection and Random Tree to detect network intrusions [5].

Himani Bhavsar, et al., presented an algorithm that adapts the idea for classification problems, such as Support Vector Machine for Data Classification. It has been developed as robust tool for classification and regression in noisy, complex domains. The two key features of support vector machines are generalization theory, which leads to a principled way to choose an hypothesis and kernel functions, which introduce nonlinearity in the hypothesis space without explicitly requiring a non-linear algorithm [6].

Chih-Wei Hsu, et al., compared the classification accuracy for the decision tree methods. They calculated the performance with three methods based on binary classifications one-against-all," "one-against-one," and directed acyclic graph SVM (DAGSVM). Their experiments indicate that the "one-against-one" and DAG methods are more suitable for practical use than the other methods [7].

III. METHODOLOGY

Using the various classification algorithms we find the best algorithm for the breast cancer dataset. The flow diagram for the comparative analysis is show in fig1:

A. Data Set

The breast cancer dataset has been collected from the UCI Repository database. This dataset contains 287 instances and 10 attributes. The data mining tool rapid miner is used for analyzing the performance of these classification algorithms.

B. Classification

In this paper we have analyzed the classification algorithms to predict which of the algorithm is most suitable for the breast cancer dataset. Classification models are tested by comparing the predicted values to known target values in a set of test data. The historical data for a classification project is typically divided into two data sets: one for building the model; the other for testing the model. In these classifications we compare three algorithms namely Naïve Bayes, Random Tree and Support Vector Machine to find out which one fits effectively for the breast cancer dataset.

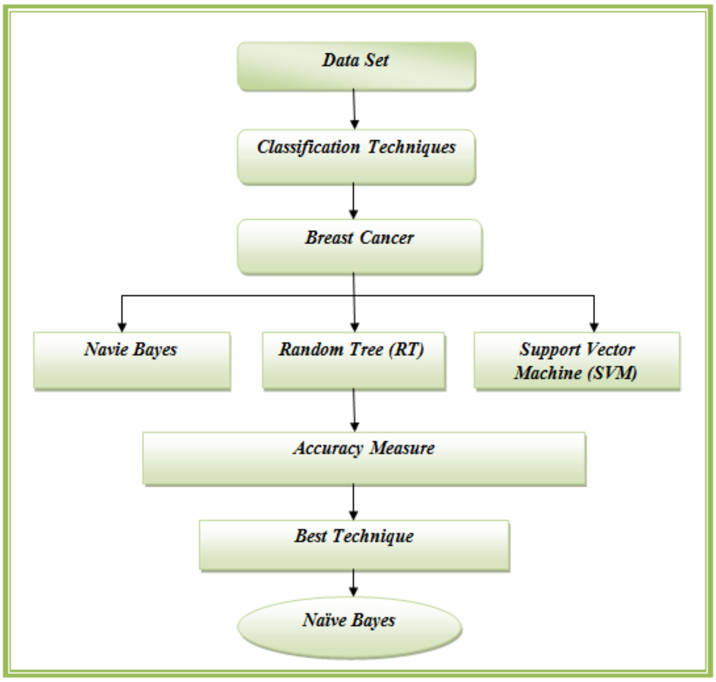

Fig 1: Flow diagram for comparative analysis

The classification algorithms are listed below.

1. Naïve Bayes
2. Random Tree (RT)
3. Support Vector Machine (SVM)

1. Navie Bayes

In classification, Bayes' rule is used to calculate the probabilities of the classes and it is a big issue how to classify raw data rationally to minimize expected risk. Probability theory is the framework for making decision under uncertainty. Bayesian theory can roughly be boiled down to one principle: to see the future, one must look at the past. Naive Bayes classifier is one of the mostly used practical Bayesian learning methods [8].

2. Random Tree

The random tree is constructed randomly from a set of possible trees having *K* random features at each node. "At random" in this context means that, in the set of trees each tree has an equal chance of being sampled or those trees have a "uniform" distribution. Random trees can be generated efficiently and the combination of large sets of random trees generally leads to accurate models. There has been an extensive research in the recent years over Random trees in the field of machine learning [9].

3. Support Vector Machine

The support vector machine (SVM) solves a quadratic programming (QP) problem in a number of coefficients equal to the number of training examples. For very large datasets, standard numeric techniques for QP become infeasible. Practical techniques decompose the problem into manageable sub-problems over part of the data or, in the limit, perform iterative pair-wise or component-wise optimization. An SVM "incrementally" on new data by discarding all previous data except their support vectors, gives only approximate results [10].

IV. EXPERIMENTAL MEASURES

In this section, we concentrate on the classification performance of the Navie Bayes, Random Tree and the Support Vector Machine for classification accuracy. And also we find out the accuracy measure and error rate to determine the best algorithm for the breast cancer dataset. It will also help in the selection of appropriate model. The accuracy measures for classification algorithms are compared with correctly classified instances and incorrectly classified instances, while the performance measures are compared with Precision Kappa and Recall.

The accuracy measure by class for the classifier algorithms is depicted in Table 1, performance measures for these classification algorithms are listed in Table 2 and the error rate measures for the classification algorithms are listed in Table 3.

Table 1: Accuracy Measures for Classification algorithms

Algorithm	Correctly classified instances (% value)	Incorrectly classified instances (% value)
Naïve Bayes (NB)	71.84	28.16
Support Vector Machine (SVM)	70.30	29.70
Random Tree (RT)	70.30	29.70

From the experimental results in Table 1, it is inferred that for Naïve Bayes algorithm, the accuracy measure is higher than the SVM and Random Tree classification algorithms for breast cancer dataset using cross validation parameter. The comparison of accuracy measures for classification algorithms are shown in Fig 2. From the Table 2, it is inferred that for the Naïve Bayes algorithm, the Precision, Kappa and Recall values are higher than the SVM and Random Tree classification algorithms. The comparison of performance measures for classification algorithms are shown in Fig 3.

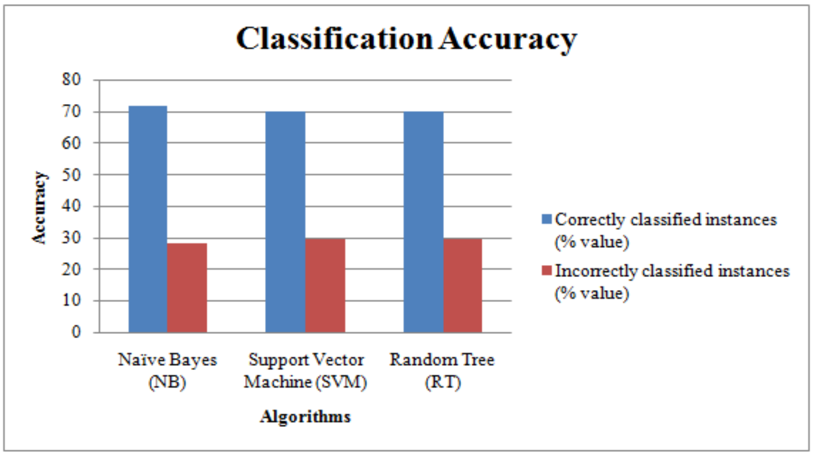

Fig 2: Comparison of Accuracy Measure for Classification algorithms

Table 2: Performance Measures for Classification algorithms

Algorithm	Precision (% Value)	Recall (% Value)	Kappa
Naïve Bayes (NB)	65.04	63.84	0.283
Support Vector Machine (SVM)	35.15	50	0
Random Tree (RT)	35.15	50	0

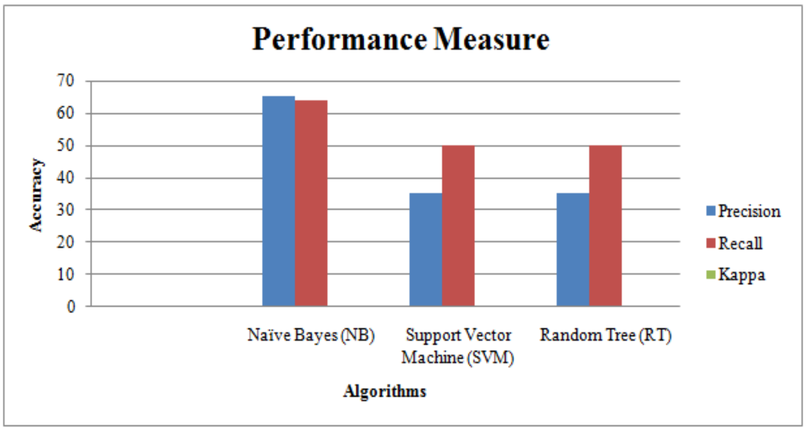

Fig 3: Comparison of Performance Measure for Classification algorithms

From the Table 3, it is inferred that for Naïve Bayes algorithm, the error rate values such as MAE and MRE are lower than the SVM and Random Tree classification algorithm for Breast cancer dataset using cross validation parameter. And for the RMSE, and RRSE values the Naïve Bayes algorithm provides higher value than the other algorithms. The comparison of error rate measures for classification algorithms are shown in Fig 4.

Table 3: Error rate Measures for Classification algorithms

Algorithm	Mean Absolute Error (MAE)	Mean Relative Error (MRE) % value	Root Mean Squared Error (RMSE)	Root relative squared error (RRSE)
Naïve Bayes (NB)	0.334	33.45	0.462	1.559
Support Vector Machine (SVM)	0.418	41.77	0.457	1.541
Random Tree (RT)	0.418	41.77	0.457	1.541

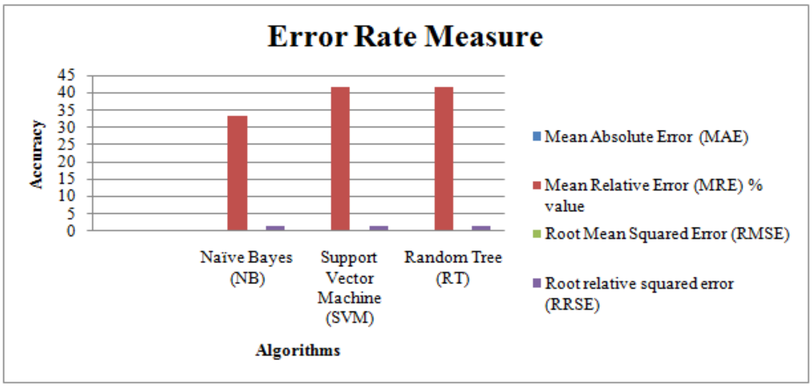

Fig 4: Comparison of Error Rate Measures for Classification algorithms

V. CONCLUSION AND FUTURE WORK

In this paper we are analyzed the performance of 3 classifiers Naive Bayes, Random Tree, Support Vector Machine. We used the breast cancer dataset for calculating the performance of those algorithms. And finally we analyzed the algorithms by using the performance factors such as the classification accuracy and error rates. From the results, it is observed that the Naïve Bayes algorithm performs better than other algorithms. In future the Naïve Bayes classification algorithm can be experimented on other datasets also. And in future we can modify the algorithm to obtain more effective results.

REFERENCES

[1] Ramyachitra and P.Manikandan, “Data mining techniques for protein sequence analysis”, Lambert Academic Publishing, 978-3-659-54129-2.

[2] Ron Kohavi, “Scaling Up the Accuracy of Naive-Bayes Classifiers a Decision-Tree Hybrid”, Silicon Graphics,Inc.Mountain View, CA 94043-1389.

[3] Jyoti Soni, Ujma Ansari, Dipesh Sharma, Sunita Soni,” Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction”, *International Journal of Computer Applications (0975 – 8887) Volume 17– No.8, March 2011.*

[4] A.Jameela, P.Revathy, ”Comparisom of Decision and Random Tree Algorithms on A Web Log Data for Finding Frequent Patterns”, *IJRET: International Journal of Research in Engineering and Technology eISSN: 2319-1163 | pISSN: 2321-7308.*

[5] Virendra Raghuwanshi, Mahendra Singh Sisodia,” Anomaly Base Network Intrusion Detection by Using Random Decision Tree and Random Projection A Fast Network Intrusion Detection Technique”, *Network Protocols and Algorithms ISSN 1943-3581 2011, Vol. 3, No. 4*

[6] Jehad Ali, Rehanullah Khan, Nasir Ahmad, Imran Maqsood,” Random Forests and Decision Trees”, *IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 3, September 2012 ISSN (Online): 1694-0814 www.IJCSI.org.*

[7] Chih-Wei Hsu, Chih-Jen Lin,” A comparison of methods for multiclass support vector machines”, Neural Networks, IEEE Transactions on (Volume:13 , Issue: 2), Page(s):415 – 425 ISSN :1045-9227.

[8] Islam, M.J, Wu, Q.M.J. , Ahmadi, M. , Sid-Ahmed, M.A.,” Investigating the Performance of Naive-Bayes Classifiers and K- Nearest Neighbor Classifiers”, Convergence Information Technology, 2007. International Conference on Date of Conference: 21-23 Nov.2007 Page(s: 1541 – 1546, ISBN: 0-7695-3038-9, IEEE.

[9] A.Jameela, P.Revathy,” Comparison of Decision and Random Tree Algorithms on A Web Log Data for Finding Frequent Patterns”, IJRET: International Journal of Research in Engineering and Technology eISSN: 2319-1163 | pISSN: 2321-7308

[10] G Cauwenberghs and T. Poggio.” Incremental and Decremental Support Vector Machine Learning”, In Advances in Neural Information Processing Systems, volume 13, 2001.

