An Efficient Bayes Classifiers Algorithm on 10-fold Cross Validation for Heart Disease Dataset

R.Nithya Dr.D.RamyachitraP.Manikandan

M.philResearch Scholar Assistant Professor Ph.DResearch Scholar
Department of Computer Science, Department of Computer Science, Department of Computer Science,
Bharathiar University, Bharathiar University, Bharathiar University,
Coimbatore, Coimbatore, Coimbatore,
Tamil Nadu. Tamil Nadu. Tamil Nadu.

***Abstract*The Classification technique forecast the categorical and prediction models to predict continuous valued functions. Generally, classification is the process of organizing data into categories for its most effective and capable use. The data classification method makes essential data that is easy to find and retrieve. In this paper the performance of three Bayes classifiers algorithms namely Naïve Bayes, Bayes Net and Naïve Bayes Multinomialare analyzed. The heart disease dataset is used for estimating the performance of the algorithms by using the cross validation parameter. And finally the comparative analysis is performed by using the factors such as the classification accuracy and error rates on all algorithms.**

Keywords - Bayes classification, Bayes Net, Naïve Bayes Multinomial Text, Naïve Bayes, Cross Validation, Heart disease dataset.

I. IINTRODUCTION

The Classification technique is a significant data mining technique with large applications. And it is used to categorize each item in a set of data into one of predefined set of groups or classes. The Classification algorithm plays a vital role in document classification. The aim of the classification technique is to construct a model in training dataset to predict the class of future objects whose class is not identified. The aim of classification is to properly forecast the assessment of a designated discrete class variable, given a vector of attributes. Classification is a classic data mining technique based on machine learning. Mainly, classification is used to classify each item in a set of data into one of predefined set of classes or groups [1].

In this paper comparison is made to find out which test option is the best for Bayes classifiers algorithm called Bayes Net, Naïve Bayes, and Naïve Bayes Multinomial. In the test option there are four kinds of parameter like supplied test set, training set, percentage spilt and cross validation. The cross validation parameter is used to calculate the data set values. This paper uses the Heart Disease dataset for comparison of those algorithms. And the paper is organized as follows. Section 2 describes the literature review, Section 3 describes the methodology for the heart disease dataset and Section 4 describes the experimental result. And finally Section 5 gives the Conclusion and Future work.

II. LITERATURE REVIEW

Waleed Ali, et al., proposed a Naïve Bayes (NB) classifier that is used to enhance the performance of conventional web proxy catching approaches such as Least-Recently-Used (LRU) and Greedy-Dual-Size (GDS). The Naïve Bayes is intelligently incorporated with conventional Web proxy caching techniques to form intelligent and effective caching approaches known as NB-LRU, NB-DA and NB-GDS. Their experimental results had revealed that their proposed NB-LRU, NB-GDS and NB-DA significantly improve the performance of the existing web proxy caching approaches across several proxy datasets[2].

Eunseog Youn, et al., developed a new feature scaling method, described class-dependent-feature-weighting (CDFW) using Naïve Bayes (NB) classifier. A new feature scaling technique, CDFW-NB-RFE, combines CDFW and recursive feature elimination (RFE). Their experimental results showed that CDFW-NB-RFE outperformed other popular feature ranking schemes used on text datasets [3].

Pablo Bermejo et al., deals with the problem of wrapper feature subset selection (FSS) and developed a proposal that is based on the combination of Naïve Bayes with incremental wrapper FSS algorithms. The merit of their approach is analyzed both theoretically and experimentally, and the results show an impressive speed-up for the embedded FSS process [4].

Luis M. de Campos, et al., proposed a new algorithm for learning Bayes Nets based on a recently introduced metaheuristic, which has survived effectively applied to solve a variety of combinatorial optimization problems like Ant Colony Optimization (ACO). They describe all the elements necessary to tackle their learning problem using this metaheuristic, and experimentally evaluate the concert of their ACO-based algorithm with other algorithms used in their literature.Their experimental work is carried out using three different domains namely ALARM, INSURANCE and BOBLO [5].

Junzhong Ji, et al., proposed a hybrid method to discover the knowledge represented in Bayesian Networks. The hybrid technique combines dependency investigation, ant colony optimization (ACO), and the simulated annealing strategy. In the first step, the new method uses order-0 independence tests with a self-adjusting threshold value to reduce the size of the explore space, hence that the search process takes less time to find the near-optimal solution. In the Second step enhanced Bayesian Network models are generated by using an improved ACO algorithm, wherever a new heuristic function is established to further enhance the search effectiveness. In the Final step, an optimization scheme based on simulated annealing is employed to improve the optimization efficiency in the stochastic search process of ants. In a number of experiments and comparisons, the hybrid technique does better than the novel ACO-B which uses ACO and some other network learning algorithms [6].

V. Muralidharan et al., presented the use of Naïve Bayes algorithm and Bayes Net algorithm for fault diagnosis through discrete wavelet features extracted from vibration signals of good and faulty conditions of the components of centrifugal pump. Classification accuracies of unusual discrete wavelet families were calculated and compared to find the best wavelet for the fault diagnosis of the centrifugal pump [7].

Pablo Bermejo et al., identified the imbalance among classes/folders as the main problem, and proposed a new method based on learning and sampling probability distributions. Their experiments over a standard corpus (ENRON) with seven datasets (e-mail users) show that the results obtained by Naïve Bayes Multinomial significantly improve when applying the balancing algorithm first. For the sake of completeness in their experimental study and also they compare this with another standard balancing method (SMOTE) and classifiers [8].

Ashraf M. Kibriya et al., presented empirical results for several versions of the Multinomial Naïve Bayes classifiers on four text categorization problem, and a way of improving it using locally weighted learning. More explicitly, it compares standard multinomial Naïve Bayes to the recently proposed Transformed Weight-Normalized Complement Naïve Bayes classifier (TWCNB), and shows that some of the modifications included in TWCNB may not be necessary to achieve optimum performance on some datasets. Finally, it shows how the performance of Multinomial Naïve Bayes can be improved using locally weighted learning [9].

Kibriya, Ashraf Masood, et al., presented empirical results for several versions of the Multinomial Naïve Bayes classifiers on four text classification problems, and a way of improving it using locally weighted learning. And finally, it shows how the performance of Multinomial Naïve Bayes can be improved using locally weighted learning [10].

III. METHODOLOGY

By using the Bayes classification technique we find the best algorithm for the heart disease dataset based on the cross validation parameter. The flow diagram for the comparative analysis is shown in Figure 1.

3.1 Dataset

Heart disease comes under the class of cardiovascular disease. The Cardiovascular disease refers to any disease that affects the cardiovascular system. The causes of cardiovascular disease are miscellaneous but atherosclerosis and hypertension are the majority one. Besides, with aging comes an amount of physiological and morphological changes that vary cardiovascular function and lead to increased risk of cardiovascular [16]. The heart disease datasets has been collected from the keel repository. This dataset contains 271 instance and 14 attributes. The data mining tool weka is used for analyzing the performance of the Bayes classification algorithm.

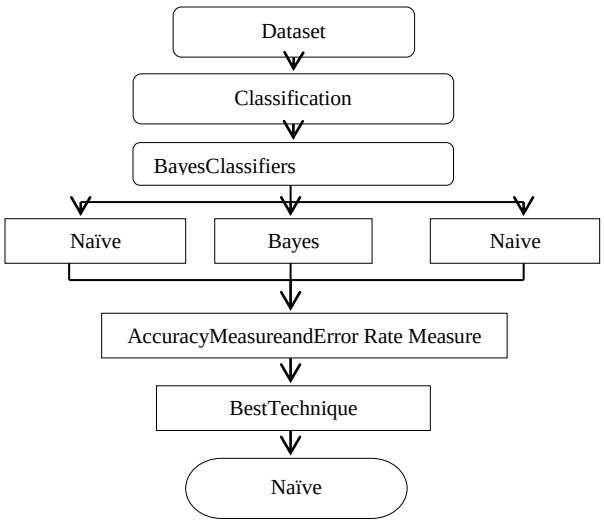

Figure 1: Flow diagram for comparative analysis of Bayes classification technique

3.2 Classification

The classification methods cluster the data into the classes on the source of their differences. A number of the classification techniques or classifiers are the Neural Network Classifier, Naïve Bayes Classifier and so on.

Every one of the method make use of the learning algorithm that generates the model that best fits the relationship between the predictors and the prediction [11]. In this paper the Bayes Classifiers algorithms are analyzed to predict which of the algorithm is most suitable for the Heart Disease dataset. In the Bayes classification technique three algorithms are compared namely Naïve Bayes, Bayes Net and Naïve Bayes Multinomial to find out which one fits effectively for the Heart Disease dataset.

3.3 Bayes classifiers

Bayes is one of the classification techniques. In this paper three bayes classification algorithms are used for finding the best algorithm for the Heart disease dataset and they are as follows.

1. Naïve Bayes

2. Bayes Net

3. Naïve Bayes Multinomial

3.3.1 Naïve Bayes

The Naive Bayes Classification technique is based on Bayesian theorem. The Naive Bayes classifiers are very scalable, involving a number of parameters linear in the number of variables in a learning problem. The Maximum-likelihood training can be completed by evaluating a closed-form expression, which takes linear time, rather than by expensive [12].

Pseudo code for Naïve Bayes:

The Naïve Bayes classifier selects the most likely classification Vnbgiven the attribute values a1, a2 …an. This results in:

$$Vnb = \text{argmax} v_j \in V\, P(v_j) \prod P(a_i \mid v_j) \quad (1)$$

Estimate P (ai | vj) using m-estimates:

$$P(ai \mid vj) = \frac{n_c + mp}{n+m} \quad (2)$$

Where:

n= the number of training examples for which v = vj

nc= number of examples for which v=vj and a=ai

p= a priori estimate for P (ai | vj).

m= the equivalent sample size.

3.3.2 Bayes Net

By means of using the bayes theorem Bayes Net can be developed. To make up a Bayesian network first conditional probability of each node must be calculated. The Acyclic graphs are used to characterize the network. Earlier than building thenetwork, it is understood that there are no missing values and all attribute values are nominal. Special types of estimators (BayesNetEstimator …) and algorithms (Hillclimber …) were used to approximate the probability. The output was visualized by using graph [13].

Pseudo code for BayesNet [17] :

1. E ⟵ Ø
2. T ⟵ *probability Tables* (*E, D*)

3. B ← (*u, E, T*)
4. *Score* ← - ∞
5. do:
   (a) *maxscore* ← *score*
   (b) for each attribute pair (x, y) do
   (c) for each $E' \in \{E \cup \{X \rightarrow Y\},$

$$E - \{X \rightarrow Y\},$$

$$E - \{X \rightarrow Y\} \cup \{Y \rightarrow X\}\}$$

   (d) T' ← *ProbabilityTables* (*E', D*)
   (e) B' ← (*u, E', T'*)
   (f) *newscore* ← *BICscore (B', D)*
   (g) if *newscore>score* then

$$B \leftarrow B'$$

$$Score \leftarrow newscore$$

6. While *score>maxscore*
7. Return *B*

3.2.3 Naïve Bayes Mutinomial

The Multinomial event model referred to as Multinomial Naive Bayes (MNB) commonly outperforms the multivariate one [14] and it also initiate to compare favorably with more specialized event models [15]. For tackling the text classification problems Naive Bayes Multinomial text algorithm is used.

Pseudo code for NaiveBayesMultinomial：

TRAINMULTINOMIALNB(C, D)

1. **V** ← EXTRACTVOCABULARY(CD)
2. **N** ← COUNTDOCS(D)
3. for each $_c$ € C
4. do N_C ← **C**OUNT**D**OCS**IN**C**L**ASS(D,$_c$)
5. *Prior[c]* ← N_c / N
6. $text_c$ ← CONCATENATE**T**EXT**O**F**ALL**DOCS**IN**C**L**ASS(**D**,$_c$)
7. **for each** t € V
8. Do T_{ct} ← COUNTTOCKENSOFTERM(text$_c$, t)
9. **do** ***condprob*[t][c]** ← $\frac{T_{ct+1}}{\sum t'(Tct'+1)}$
10. return v, prior, condprob

APPLY **M**ULTINOMIAL
NB(C,V,*prior*,*condprob*,*d*)
1. w ← EXTRACT**T**OKENS**F**ROM**D**OC(V,D)
2. **for each** c € C
3. **do** score[c] ← log prior[c]
4. **for each** t € w
5. **do** score [c]+ = log condprob[t][c]
6. return argmax $_{c \in C}$ score[c]

Figure 4: Pseudo code for CART classification algorithm

IV. EXPERIMENTAL RESULTS:

In this paper the experimental measures is calculated by using the performance factors such as the classification accuracy and error rates. And also we find out the comparative analysis for the heart disease dataset to predict the finest algorithm. The accuracy measure and the performance factors by class for the Bayes classifiers is depicted in Table 1.

Table 1: Comparison of performance factors for Bayes classifiers algorithms

Algorithm	Correctly classified instances (% value)	Incorrectly classified instances (% value)
NaïveBayes	83.7037%	16.2963%
BayesNet	83.3333%	16.6667%
NaiveBayesMultinomial	73.7037%	26.2963%

From the Table 1, it is inferred that for Naïve Bayes Algorithm on cross validation parameter, the TP rate, Precision, F-Measure, ROC curve and the Kappa Values are higher than the other two algorithms such as the Bayes Net and Naïve Bayes Multinomial. The comparisons of performance measures for Bayes classifier has shown in Figure6and the accuracy measure for the bayes classifiers is shown in Table 2.

Algorithms	TP Rate	Precision	F-Measure	ROC Curve	Kappa value
NaïveBayes	0.837	0.837	0.837	0.837	0.6689
BayesNet	0.833	0.833	0.833	0.833	0.6617
NaiveBayesMultinomial	0.737	0.736	0.736	0.799	0.4635

Table 2: Comparison of accuracy measure for Bayes classifiers algorithms

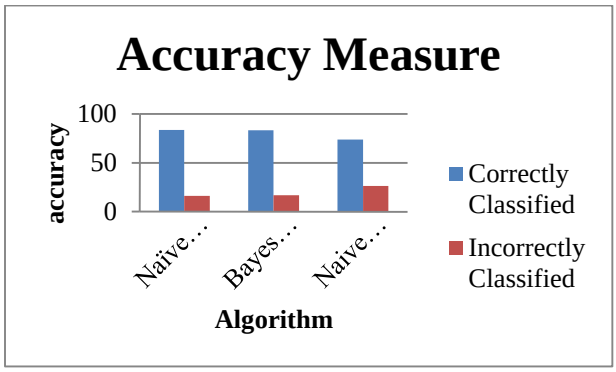

Figure 5:Comparison of Accuracy Measure for Classification algorithms

From the Table 2, it is inferred that the Naïve bayes algorithm has higher classification accuracy than the other classification algorithms such as the Bayes Net and Naïve Bayes Multinomial. The comparison of the accuracy measures for the bayes classifiers is shown in Figure5and the error rate measures for the bayes classifiers are shown in Table 3.

For Correctly Classified instances, it is inferred that Naïve Bayes algorithm performs 0.44% better than BayesNet and 11.94% better than Naïve Bayes Multinomial. Similarly for incorrectly classified instances it is inferred that Naïve Bayes algorithm performs 2.22% better than BayesNet and 38.02% better than Naïve Bayes Multinomial.

For TP rate, it is inferred that Naïve Bayes algorithm performs 0.47% better than BayesNet and 11.94% better than Naïve Bayes Multinomial for TP rate. For precision it is inferred that Naïve Bayes algorithm perform 0.47% better than BayesNet and 12.06% better than Naïve Bayes Multinomial. For F-measure it is inferred that Naïve Bayes algorithm is 0.47% better than BayesNet and 12.06% better than Naïve Bayes Multinomial. For ROC Curve it is inferred that Naïve Bayes algorithm is 0.47% better than BayesNet and 4.54% better than Naïve Bayes Multinomial. For kappa Value it is inferred that Naïve Bayes algorithm is 1.07% better than BayesNet and 30.70% better than Naïve Bayes Multinomial.

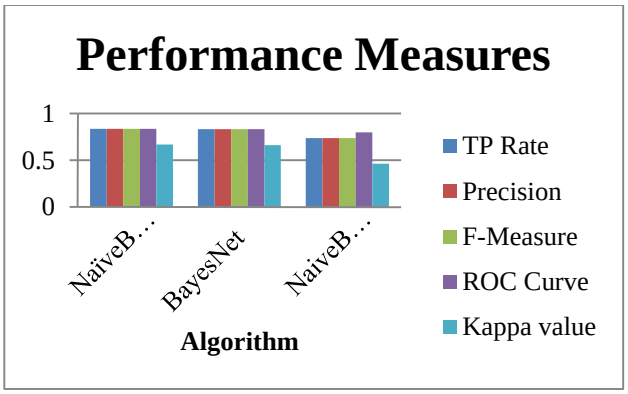

Figure 6:Comparison of performance factors for Bayes classifiers algorithms

From the Table 3, it is inferred that the Naïve Bayes classification algorithm has the lowest error rates than the other classification algorithms such as the Bayes Net and Naïve Bayes Multinomial. The comparison of the error measures for the bayes classifiers is shown in Figureures 7 and 8.

Algorithms	MAE	RMSE	RAE	RRSE
Naïve Bayes	0.1863	0.3607	37.7196	72.5867
Bayes Net	0.1947	0.3604	39.4261	72.5214
NaiveBayesMultinomial	0.2664	0.4869	53.9452	97.9937

Table 3: Comparison of error rate measures for Bayes classifiers algorithms

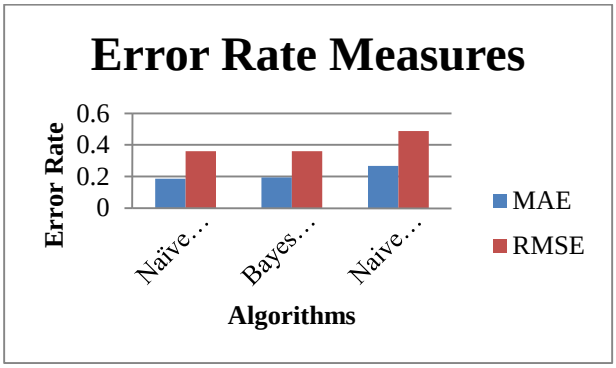

Figure 7: Comparison of error rate measures for Bayes classifiers algorithms

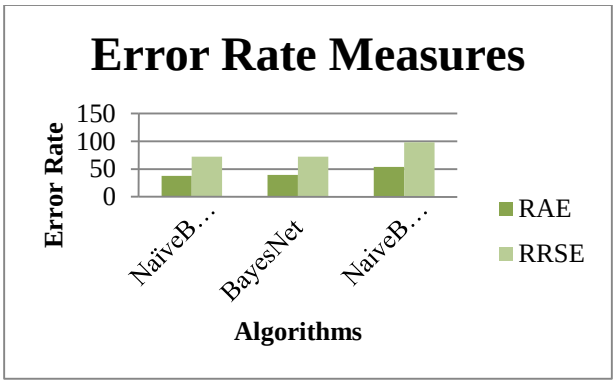

Figure 8: Comparison of error rate measures for Bayes classifiers algorithms

For MAE, it is inferred that Naïve Bayes algorithm performs 4.31% better than BayesNet and 30.06% better than Naïve Bayes Multinomial. For RMSE it is inferred that Naïve Bayes algorithm performs 0.08% better than BayesNet and 25.91% better than Naïve Bayes Multinomial. For RAE it is inferred that Naïve Bayes algorithm performs 4.32% better than BayesNet and 30.07% better than Naïve Bayes Multinomial. For RRSE it is inferred that Naïve Bayes algorithm performs 0.08% better than BayesNet and 25.92% better than Naïve Bayes Multinomial.

4. CONCLUSION:

This paper analyzed the performance of 3 Bayes classifiers algorithms namely Bayes Net, Naive Bayes, Naive Bayes Multinomial Text. The heart disease datasets is used for calculating the performance by using cross validation parameter based on the class attribute. The algorithms are analyzed based on the performance factors such as classification accuracy and error rates. From the experimental results, it is observed that the Naïve Bayes algorithm performs better than other algorithms. In the future, the classification Bayes algorithms can be experimented on other datasets to obtain more effective results. Also the Bayes classification algorithms can be analyzed by using parameters such as the training set, percentage split, and supplied test set.

References:

1. Mahendra Tiwari, et al., "Comparative Investigation Of Decision Tree Algorithms On Iris Data", International Journal of Advances in Computer Science and Technology, Volume 2, Page No: 30-35, March 2013, ISSN 2320 – 2602.
2. Waleed Ali, et al., "Intelligent Naïve Bayes-based approaches for web proxy caching", Knowledge- based system 31(2012) 162-175.
3. Eunseog Youn, et al., "Class dependent future scaling method using naïve bayes classifier for text datamining", Pattern Recognition Letters 30(2009) 477-485.
4. Pablo Bermejo et al., "Speeding up implemental wrapper feature subset selection with naïve bayes classifiers", Knowledge-based system 55(2014) 140-147.
5. Luis M. de Campos, et al., "Ant colony optimization for learning Bayesian networks", International Journal of Approximate Reasoning 31 (2002) 291–311.
6. Junzhong Ji, et al., "A hybrid method for learning Bayesian network based on ant colony optimization", Applied soft computing 11(2011) 3373-3384.
7. V. Muralidharan et al., "A Comparative study of Naïve Bayes classifier and Bayes net classifier for fault diagnosis of monoblock centrifugal pump using wavelet analysis", Applied soft computing 12(2012) 2023-2029.
8. Pablo Bermejo et al., " Improving the performance of Naïve Bayes multinomial in e-mail foldering by introducing distribution-based balance of dataset", Experts system with applications 38(2011) 2072-2080.
9. Ashraf M. Kibriya et al., "Multinomial Naïve Bayes for Text Categorization Revisited", Artificial Intelligence 2004, LNAI 3339, pp. 488–499, 2004.
10. Kibriya, Ashraf Masood, et al., "Multinomial Naïve Bayes for Text Categorization Revisited",http://hdl.handle.net/10289/1448.
11. Shabia Shabir Khan et al., "Evaluation of Knowledge Extraction Using Various Classification Data Mining Techniques", International Journal of Advanced Research in Computer Science and Software Engineering , Volume 3,pp. 251-256, June 2013, ISSN: 2277 128X.
12. http://en.wikipedia.org/wiki/Naive_Bayes_classifier#The_naive_Bayes_probabilistic_model
13. Samir Kumar Sarangi and Dr. Vivek Jaglan, "Performance Comparison of Machine Learning Algorithms on Integration of Clustering and Classification Techniques", International Journal of Emerging Technologies in Computational and Applied Sciences(IJETCAS), IJETCAS 13-144, ISSN:2279-0047.
14. McCallum, A., Nigam.K, "A comparison of event models for naive bayes text classification", Technical report, American Association for Artificial Intelligence

Workshop on Learning for text Categorization (1998).

15. Eyheramendy, S., Lewis, D.D., Madigan, "On the naive Bayes Model for text categorization", In: Proceedings of the Tenth European Conference on Artificial Intelligence and Statistics (2003), 3-6.

16. http://en.wikipedia.org/wiki/Cardiovascular_disease.

17. http://nlp.stanford.edu/IR-book/html/htmledition/naive-bayes-text-classification.