Indian Journal of Science and Technology, *Vol 9(40), DOI: 10.17485/ijst/2016/v9i40/99425, October 2016*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Performance Evaluation of Classification Algorithms on Different Data Sets

Meenu Gupta* **and Deepak Dahiya**

Ansal University Gurgaon, Gurgaon - 122003, Haryana, India; gupta.meenu5@gmail.com, deepakdahiya@ansaluniversity.edu.in

Abstract

Objectives: The most appropriate classifier selections for the particular data sets were generally found harder. Therefore, in this study various existing classifiers have been considered on several data sets to assess their performance. **Methods/ Statistical Analysis:** Usually, the selections of classification techniques, such as, Naive Bayes (NB), Decision Tree (DT), Lazy Classifiers (LC), Support Vector Machine, etc., depend on the type and nature of the attributes in the data set. The wrong selection of classification technique can certainly lead to wrong results and poor performance. This concept is the motivation behind this study. Usually the data set consists of nominal attributes, numeric attributes or mix attributes (both numeric and nominal attribute). In this paper, different types of data sets are applied on three most popular classification techniques, such as, NB, DT, and LC, to evaluate their performances. **Findings:** The result reveals that NB classifier performs well on both mix attribute data and numeric data but decision tree classifier performs better on nominal attribute data. Lazy classifier's performance is just average for all kind of data. **Application/Improvements:** The results of this study will helps in understanding the performance of different classification techniques on different data sets. Further, results can be utilized to select the best classification technique among NB, decision tree and lazy classifiers in order to use with different data sets.

Keywords: Accuracy, Classification, Data set, Decision tree, Lazy Classifiers, NB

1. Introduction

Data mining is the process to extract potentially valuable and relevant information from big amount of data sets[1]. Usually, it includes a set of technique, such as, classification, clustering, association rule mining, anomaly detection, etc[2]. Data mining techniques have been widely used to analyze data from different domains such as business[3,4], medical[5,6] and transportation[7–12]. The wide applicability of these data mining techniques proved it as a reliable and result oriented in all real world domains. Classification or prediction is the most commonly used techniques of data mining. Classification is a kind of supervised learning techniques that identifies the hidden relationships between dependent and independent variables[13]. Supervised learning techniques extract certain important features from the training data and then it uses those features to test on unobserved data[2]. A wide application of classification techniques is image classification, pattern recognition, medical disease diagnosis, fault detection, traffic accident severity analysis and detecting financial trends[14].

In order to use the classification model for actual implementation, certain criteria are used to validate the performance of the model[1, 2]. Several types of classification techniques are existing, such as, NB, DT, LC, SVM, K-Nearest Neighbor, ANN, etc. The performance of all the classification algorithms is not similar on all data types. In other words, the performance of different classifiers is varied on different data sets. The data sets can have three basic types of attribute values: numeric, nominal or both. Therefore, the selection of any classification algorithms must utilize the knowledge about data and its attribute values. Wrong selection of classification algorithm will

*Author for correspondence

certainly lead to bad classification model and bad results. This motivates our study.

This paper evaluates the performance of most popular classification algorithms, namely, NB, DT and LC on three different types of data sets. The outcome of this study will certainly contributes in identifying if the different characteristics of the data affect the performance of classifiers. Also, we will identify that for what kind of data, which classification algorithms will be more suitable.

This study would be helpful for the beginners to choose among the set of classification techniques to perform on a variety of data set. The organization of the paper is as follows: Data sets used in this study and methodology has been discussed in Section 2. Analysis of the results has been presented in Section 3 and finally the paper is concluded with future scope in Section 4.

2. Materials and Methods

The various data sets and the methodology adopted to analyze these data sets have been incorporated in this section for the discussion.

2.1 Data Description

The three different types of data set have been used for this study. Breast cancer data with all nominal attributes, diabetes data with numeric attributes and German credit card data with both numeric and nominal attributes are used. All these data sets have target attribute or dependent attribute value as nominal. The brief description of number of data instances and attributes are given in Table 1.

Table 1. Description of data set used

Data Set name	Type of data	Number of attribute	Number of instances
Breast cancer data	Nominal	10	286
Diabetes data	Numeric	9	768
German credit card data	Mix	21	1000

2.2 Classification Techniques

To build a classification rules, training data is given to classifier. This technique of supervised learning is usually termed as classification. Three popular classification techniques NB decision tree and lazy classifiers are used for this study. A description of these techniques is given as follows:

2.2.1 Naive Bayes (NB)

A NB classifier is a probabilistic framework for solving classification problems based on conditional probability and Bayes theorem. NB classifier consider one feature of a particular class is unrelated the occurrence of any other feature of a class. A fruit may be considered as a banana can be of yellow color and long even though these features depend upon each other but they contribute independently[15]. This independent assumption between predictors is known as 'naïve'. Naive Bayesian learning is found more accurate in test set than any other known method, including ANN and DT in real world data set[16]. Posterior probability, P(ci|x), from P(ci), P(x), and P(x|ci), is usually calculated through Bayes theorem. According to assumption, namely, class conditional independence[15,16], the posterior probability is given as:

$$P\left(\frac{c_i}{x}\right) = \frac{P\left(\frac{x}{c_i}\right).P(c_i)}{P(x)}$$

$P(c_i|x) = P(x_1|c_i) * P(x_2|c_i) *\ldots\ldots\ldots..* P(x_n|c_i) * P(c_i)$

Here, P(ci|x), P(ci), P(x|ci), and P(x) represent posterior probability of class, prior probability of class, predictor probability, and prior probability of predictor, respectively.

2.2.2 Decision Tree

In order to discover useful patterns, prediction of the value of large and complex bodies of categorical data is highly required. The most popular machine learning technique used to perform this task DT. Decision tree is a predictive model which has root node, leaf node and branches of a large data set that maps observation about target value as a conclusion[21–23]. Decision tree follows top down strategy for implementation on a large set of data without losing any information. Decision tree prevent from over fitting and handling of missing data that can be leave by other technique as a burden to the user. Decision tree is useful in segmentation of data, analyze the effect of changing one variable to another, data processing and prediction of a variables in a data set that can eventually be used as a target. A decision tree algorithm analyses the data and creates a repeating series of branches until no

more relevant branches can be made. The end result is a binary tree structure where the splits in the branches can be followed along specific criteria to find the most desired result. Decision tree used Gain to represent the difference between the amount of information that is needed to correctly make a prediction both before and after the split has been made. ID3 algorithm uses Entropy and Information Gain to construct a decision tree. The entropy calculates on the basis of homogeneity of a sample data. Zero entropy represents homogeneous and unity represents equally divided.

$$\text{Entropy (T)} = \sum_{i=1}^{c} -p_i \log_2 p_i \quad (1)$$

For example, if we have 4 +ve sample and 3 –ve sample in a node then the estimated probability of +ve is 4/7 = 0.57118. Information gain, based on decrease in entropy, is the other important part of constructing a decision tree. Usually, construction of the decision tree indicates attribute that reflects the highest information gain (i.e., the most homogeneous branches).

$$Gain\,(C, X) = Entropy\,(C) - Entropy\,(C, X) \quad (2)$$

Where C is the target class and x is the attribute.

2.2.3 Lazy classifiers

Instance based learning which simply stores training data and wait until a query is made to the system where the system tries to generalize the training data before receiving queries is LC. Local target function for each query system is the major advantage of LC. The main drawback of LC is the requirement of large apace to save the training data and slower performance[19–25].

2.3 K-Fold Cross Validation

The outcomes of a statistical analysis will simplify to an independent data set through the model techniques, namely, cross validation[26]. Among the all available cross validation techniques, K-fold cross validation is always considered as the most common technique for the estimation of the performance of a classifier. Single run of k-fold cross validation, from the set of m training examples, can be estimated in the following steps:

- Firstly, training data arranged in random fashions.
- Then, training data sets are distributed in k folds.
- For i = 1. . . k:
  - The classifiers, not belonging to Fold i, have been trained.
  - Test the classifier of Fold i on all the examples.
  - Compute n_i, the number of examples in Fold i that were wrongly classified.
- Return the following estimate to the classifier error:

$$E = \frac{\sum_{i=1}^{k} n_i}{m} \quad (3)$$

To achieve more accuracy of a classifier, k-fold cross validation is run several times, each with a different random arrangement in Step 1. Let E1, . . . , Et be the accuracy estimates obtained in t runs. Define:

$$e = \frac{\sum_{j=1}^{t} E_j}{t}, \qquad V = \frac{\sum_{j=1}^{t} (E_j - e)^2}{t-1}, \qquad \sigma = \sqrt{V} \quad (4)$$

The estimate for the algorithm performance is an error of e with standard-deviation of σ[21].

2.4 Accuracy Measures

Different accuracy measures that are used to evaluate the performance of classifiers in this study are described below:

2.4.1 Recall and Precision

In the recall and precision we find out either our data is relevant or not. Recall in information retrieval is the fraction of the documents that are relevant to the query that are successfully retrieved.

$$Recall = \frac{t_p}{t_p + f_n} \quad (5)$$

Precision in the information retrieval is the fraction of retrieved documents that are relevant to the query.

$$Precision = \frac{t_p}{t_p + f_p} \quad (6)$$

Whereas t_p, f_p, f_n are the true positive, false positive and false negative. There is a tradeoff between Recall and precision, if recall increases then precision decreases vice versa. There is a problem with calculating precision and recall is to be considered the record must be relevant or irrelevant. The record can be completely relevant, completely irrelevant or somewhat irrelevant. This problem is arising by individual perception because it varies from person to person. Now difficulty arises in finding the

relevant record from the data base. There are many different ways to create a pool of relevant records: One is to use search method for all relevant records, another is to manually scan the entire document to identify the set of relevant records[22].

2.4.2 F-measure and ROC

F-score or measure is often defined as the ratio between square of geometric mean to arithmetic mean of precision and recall, as given in Equation (7). Bias, as evaluation metric, results in criticism of F-score[23,24,27,28].

$$F\ Score = 2 * \left(\frac{Precision * Recall}{Precision + Recall}\right) \quad (7)$$

Further, the swap between True Positive Rate (TPR) and False Positive Rate (FPR) of classifier through graphical approach is termed as Receiver Operating Characteristics (ROC).

Performance of each classifier is represented as a point on the ROC curve where TPR on y axis and FPR on x axis.

3. Results and Discussion

We performed classification on three different types of data i.e., nominal data, numeric data and mix data with both numeric and nominal attributes using NB, decision tree and lazy classifiers. The Table 2 gives the accuracy of the correct prediction of the class values on different data sets. Table 2 shows that on numeric data, NB classifier achieved the highest accuracy of 76.30% while lazy classifier achieved the lowest accuracy of 69.14%. For nominal data, decision tree obtained the highest accuracy of 75.52% while NB achieved the lowest 71.68%. For mix data with both numeric and nominal attributes, table shows that the accuracy of NB is highest with 75.4% and lowest with 69.4% for lazy classifiers. Hence, the analysis reveals that different classification techniques have different accuracy and performance for correct prediction on different data sets.

Table 3 shows the different other classifier's performance measures for above mentioned classification techniques on different type of data sets. The results in the Table 3 again illustrate the same results as shown by Table 2. The value for different performance measures such as precision, recall, F-score and ROC shows that NB technique perform superior on numeric data set and mix attribute data sets whereas the decision tree classifiers performs better on nominal data sets. The performance of lazy classifier was on average on different data sets. The different ROC curve to illustrate the performance of NB, decision tree and lazy classifiers is given in Figure 1, 2 and 3.

Table 2. Classifier's accuracy on different data sets

Classifier/Data	Numeric Data	Nominal Data	Mix Data
NB	76.30%	71.68%	75.4%
Decision Tree (J48)	73.83%	75.52%	70.5%
Lazy Classifier (K*)	69.14%	73.42%	69.4%

Figure 1 shows the performance of different classifiers on mix attribute data set. It can be seen that ROC curve for NB classifier is the superior than others for mix data. But there is a very slight variation in the ROC curve for decision tree and lazy classifier. Figure 2 illustrate the ROC curve for three classifiers for numeric attribute data set. ROC curve for all three classifiers in Figure 2 is slightly up and down, but overall NB's performance is superior for numeric attribute data considering other factors available in Table 3. Figure 3 illustrates the performance of three classifiers on nominal data set. It is clearly revealed that the ROC area for decision tree classifier is the highest among other classifiers.

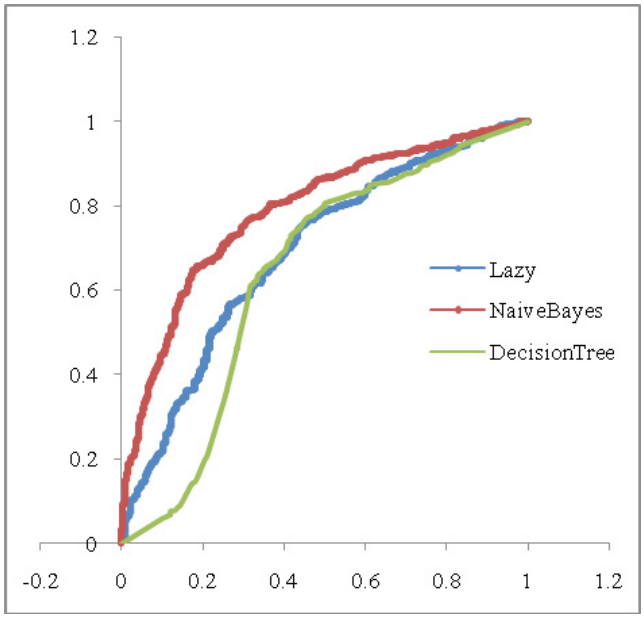

Figure 1. Performance of classifier on mix data set.

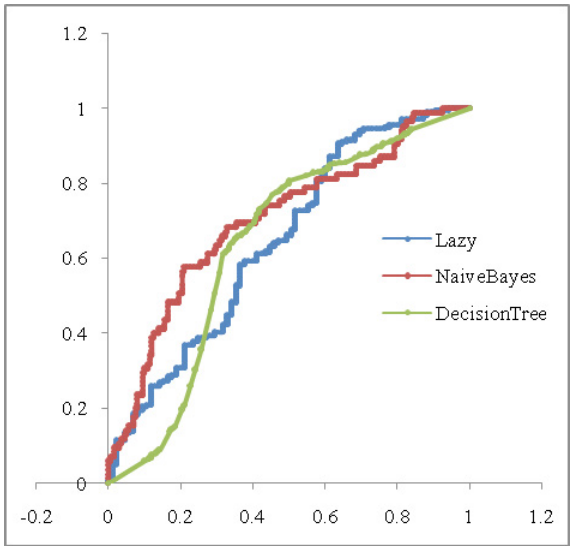

Figure 2. Performance of classifier on numeric data set.

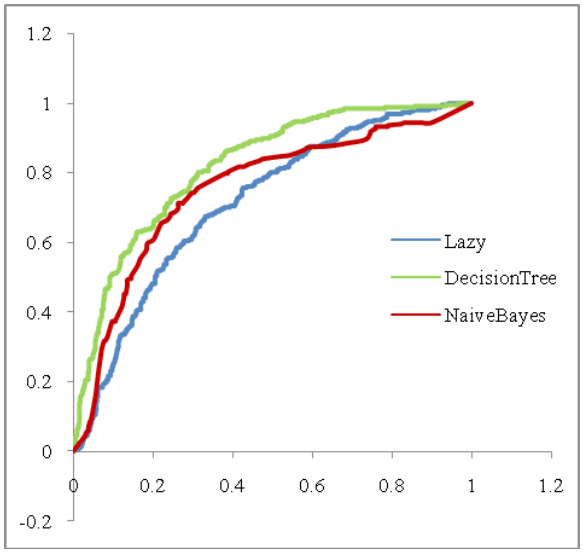

Figure 3. Performance of classifier on Nominal data set.

Table 3. Accuracy measures for different classifiers

Classifier	Data type	Precision	Recall	F-score	ROC
NB	Numeric data	0.759	0.763	0.760	0.819
	Nominal Data	0.704	0.717	0.708	0.701
	Mix data	0.743	0.754	0.746	0.787
Decision Tree	Numeric data	0.735	0.738	0.736	0.751
	Nominal Data	0.752	0.755	0.713	0.784
	Mix data	0.687	0.705	0.692	0.639
Lazy Classifier	Numeric data	0.680	0.691	0.683	0.714
	Nominal Data	0.714	0.734	0.713	0.645
	Mix data	0.682	0.694	0.687	0.689

Therefore, the experimental analysis on the performance of NB, decision tree and lazy classifiers on three different data sets i.e., numeric, nominal and mix attribute data set, illustrates that NB's performance on numeric and mix attribute data is better than decision tree and lazy classifier. Also, for nominal attribute data, our experimental analysis revealed that decision tree outperforms the NB and lazy classifiers. Hence, our study certainly helps in deciding the better classification technique based on different type of data set.

4. Conclusion and Future Work

In this study, we have used three popular classification techniques NB, decision tree and lazy classifiers on different data sets. The purpose of this study is to check the performance of different classification algorithms on different data sets. In order to achieve this, we have used three different data sets i.e., numeric data (diabetes data set), nominal data (breast cancer data set) and mix attribute data (German credit card data set). Further, we performed all three classification techniques on these data sets and compared the result. The results illustrate that the performance of NB classifier on numeric and mix data set is superior to the decision tree and lazy classifiers. The result also revealed that on nominal attribute data set, the decision tree classification technique outperformed the NB and lazy classifiers. Therefore, this study simply revealed the important information that the different classification algorithms have different accuracy and performance on different kind of data sets. Our study certainly helps the beginners to choose the best classification algorithm in order to apply different kind of data set. Our future work will consist of selection of some real world large data set and perform some suitable classification technique based on the nature and characteristics of the data and providing some important information out of the data set.

Conflict of Interest: Authors declare that they have no conflict of interest.

5. References

1. Tan PN, Steinbach M, Kumar V. Introduction to data mining. Pearson Addison-Wesley; 2006.
2. Han J, Kamber M. Data mining concepts and techniques. USA Morgan Kaufmann Publishers; 2001. p. 1–169.
3. Jukic N, Jukic B. Modeling-centered data warehousing learning Methods, Concepts and Resources. International Journal of Business Intelligence Research. 2012; 3(4):74–95.
4. Ana A, Manuel FS. Closing the gap between data mining and business users of business intelligence systems a design science approach. International Journal of Business Intelligence Research. 2012; 3(4):14–53.
5. Nauck D, Kruse R. Obtaining interpretable fuzzy classification rules from medical data. Artificial Intelligence in Medicine. 1999; 16(2):149–69.
6. Luukka P. Similarity classifier using similarity measure derived from Yu's norms in classification of medical data sets. Computers in Biology and Medicine. 2007; 37(8):1133–40.
7. Kumar S, Toshniwal D. A data mining framework to analyse road accident data. Journal of Big Data. Springer. 2015; 2(26):1–18.
8. Kumar S, Toshniwal D. A data mining approach to characterize road accident locations. Journal of Modern Transportation. Springer. 2016; 24(1):62–72.
9. Kumar S, Toshniwal D. A novel framework to analyze road accident time series data. Journal of Big Data. Springer. 2016; 3(8):1–11.
10. Kumar S, Toshniwal D. Analysis of hourly road accident counts using hierarchical clustering and cophenetic correlation coefficient. Journal of Big Data. Springer. 2016; 3(13):1–11.
11. Kumar S, Toshniwal D. Analysing road accident data using association rule mining. Proceedings in IEEE International Conference on Computing, Communication and Security held in Mauritius. India: IEEE Xplore; 2015.
12. Kumar S, Toshniwal D, Parida M. A comparative analysis of heterogeneity in road accident data using data mining techniques. Evolving Systems. Springer. 2016; 5. DOI: 10.1007/s12530-016-9165-5.
13. Maimon O, Rokach L. The data mining and knowledge discovery handbook. 2nd ed: Berlin: Springer; 2010.
14. Dunham MH. Data mining introductory and advanced topics. New Jersey: Prentice Hall; 2002.
15. Analytics vidhya. Available from: http://www.analyticsvidhya.com/blog/2015/09/naive-bayes-explained/
16. NBian. Available from: http://www.saedsayad.com/naive_bayesian.htm
17. Elkan C. Niave Baysian learning. Adapted from Technical Report No. CS97-557, Department of Computer Science and Engineering, University of California, San Diego; 1997. p. 1–11.
18. Shahrukh T, Prashasti K. A survey on decision tree based approaches in data mining. International Journal of Advanced Research in Computer Science and Software Engineering. 2015; 5(4):25–71.
19. Vijayarani S, Muthulakshmi M. Comparative analysis of bayes and lazy classification algorithms. International Journal of Advanced Research in Computer and Communication Engineering. 2013; 2(8):3118–24.
20. Durairaj M, Deepika R. Comparative analysis of classification algorithms for the prediction of leukemia cancer. International Journal of Advanced Research in Computer Science and Software Engineering. 2015; 5(8):787–91.
21. k-fold cross validation. Available from: http://www.csie.ntu.edu.tw/~b92109/course/Machine %20Learning/Cross-Validation.pdf
22. Rupali B, Sonia V. Implementation of ID3 algorithm. International Journal of Advanced Research in Computer Science and Software Engineering. 2013; 3(6):845–51.
23. Measuring search effectiveness. Available from: https://www.creighton.edu/fileadmin/user/HSL/ docs / ref/ Searching_-_Recall_Precision.pdf
24. Baeza-Yates B, Ricardo R, Ribeiro-Neto RN, Berthier B. Modern information retrieval. New York, NY: ACM Press, Addison-Wesley; 1999. p. 1–103. ISBN: 0-201-39829-X.
25. Zolfagharifar SA, Karamizadeh FV. Developing a hybrid intelligent classifier by using evolutionary learning (Genetic Algorithm and Decision Tree). Indian Journal of Science and Technology. 2016 May; 9(20):1–8.
26. Kim M, Kim CJ. Factors associated with decision to participate in physical activity by people with spinal cord injury: An analysis using decision tree. Indian Journal of Science and Technology. 2016 Jul; 9(26):1–7.
27. Azad C, Jha VK. Data mining based hybrid intrusion detection system. Indian Journal of Science and Technology. 2014 Jan; 7(6):1–9.
28. Rajalakshmi V, Mala GSA. Anonymization by data relocation using sub-clustering for privacy preserving data mining. Indian Journal of Science and Technology. 2014 Jan; 7(7):1–6.