International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 6 | ISSN : 2456-3307

Comparative Analysis of Classification Methods in R Environment with two Different Data Sets

B Nithya*1, Dr. V Ilango2

*1Senior Assistant Professor & Research Scholar, Department of MCA, New Horizon College of Engineering, Bangalore, India
2Professor, Department of MCA, New Horizon College of Engineering, Bangalore, India

ABSTRACT

Machine Learning methods are widely used in various domains as they are influential in classification and prediction processes. The frequently used supervised machine learning task is classification. There are various types of classification algorithms with strengths and weaknesses appropriate for different types of input data. This paper depicts the implementation of few classification methods such as Decision Tree, K Nearest Neighbour and Naïve Byes classifier for different datasets in R environment. This paper presents the comparative study of these methods using open source tool R. The aim of this paper is to analyse the performance of these methods in two different datasets based on the evaluation metrics like accuracy and error rate. The implementation procedure show that the performance of any classification algorithm is based on the type of attributes of datasets and their characteristics. This paper shows that based on the constraints, requirements with type of input datasets specific algorithm and tool can be chosen for implementation.

Keywords: Machine Learning, Classification, Decision Tree, K Nearest Neighbour, Naïve Bayes Classifier, Performance, R Tool.

I. INTRODUCTION

Machine Learning methods have been applied in several domains and it has been proved that their practice is inevitable in various applications *[1]*. There are various commonly used machine learning algorithms and they can be useful to almost any data problem. Based on the type of learning these algorithms can be used in real time data. In Machine Learning, Classification is the problem of identifying to which set of types or classes a new observation belongs, based on the training set of data containing observations or instances whose class membership is already known. The supervised learning algorithms in machine learning can be used for the classification and numeric predictions. The frequently used classification algorithms in machine learning are Nearest Neighbour, Decision Trees, Naïve Bayes and Rule Based classifiers. All these algorithms are having their own strengths and weaknesses based on the types of input data and tools used for implementation purpose. There are many tools available for machine learning algorithms executions and choosing exact tool can be very important for working with best algorithms. The open source tool R is one of the most powerful and popular for statistical programming and applied machine learning. Finding the efficiency of any method or tool is essential while we aim for optimization. As we have various classification algorithms and tools for machine learning, this paper aims to compare the competence of few of these classification methods by employing them in R tool with two different datasets as inputs.

II. RELATED WORK

In the past decades abundant researches have carried out using machine learning classification techniques in various fields. The performance analysis and optimization of supervised methods have been carried out on a set of data using different tools.

Prediction of the orthopaedic problems by implementing almost twenty algorithms on two different open source tools such as Weka and Tanagra *[2]* has been carried out to estimate the accuracy among all the algorithms and also the attribute ranking is developed to make a decision. This work showed that among all the classification algorithms, the results are more accurate in Tanagra tool compared to Weka.

The methods namely Bayesian network, Naïve Bayes, J48, REP, Random Forest, Random tree, CART, KNN and Conjunctive rule learning were employed on the diabetes dataset to find the best classifier for Diabetes Diagnosis *[3]*. The results showed that J48 algorithm is best for the diabetes data set.

To find the maximum accuracy with reduced subset of features the algorithms of Naïve Bayes(NB), Logistic Regression(LR) and Decision Tree(DT) classifiers were implemented on breast cancer data and the time complexity of each of the classifier also measured *[4]*. Here, Logistic Regression classifier is concluded as the best classifier with the highest accuracy as compared to other two classifiers.

For survival prediction in breast cancer dataset, few machine learning techniques were used for implementation to classify breast cancer patients using 70-gene signature *[8, 9]*. These related works exhibited that genetic programming methods are worth further investigation as a tool for cancer patient classification based on gene expression data.

For classification of clustered Micro Calcifications in digital mammograms results obtained from two different sets of experiments demonstrated that the kernel based methods Support Vector Machine (SVM), Kernel Fisher Discriminant (KFD) and Relevance Vector Machine (RVM) yielded the best performance *[10]*. Furthermore, this work verified that these methods were also computationally advantageous both in training and testing.

To predict student marks, J48, Random Forest, Naive Bayes, Naive Bayes Multinomial, K-star, algorithms were implemented on student dataset which had sixteen parameters *[6]*. This work concluded that random forest becomes more accurate with the number of entries but all algorithms need modification if they can ever be used because the current amount of accuracy is low for this to be implemented on a large scale.

An association study with the goal of finding reliable classifiers that predict the presence or absence of breast cancer genes *[7]* have compared several machine learning techniques based on the available features. Here a correlation between size and performance was noticed between the CART and J48 solutions, and among the different ClassGP solutions. This work designated that CART regression trees as the best classifiers, both in terms of performance and interpretability.

The analysis of different classifiers using WEKA tool has revealed the results based on the accuracy of classifiers and the time taken by these classifiers for classifying various sizes of datasets *[5]*. This work has concluded that each algorithm has its own set of advantages and drawbacks as well as its own area of implementation. Also showed that none of the algorithm can satisfy all the constrains and criteria, hence it is based on the applications and requirements, specific algorithm can be chosen.

Based on these studies, few of the Machine Learning classification methods are considered in this work with two datasets and they are implemented in R environment.

III. MACHINE LEARNING CLASSIFICATION ALGORITHMS

Classification is one of the most extensively used techniques in machine learning, with a wide-ranging array of applications like sentiment analysis, ad targeting, spam detection, risk assessment, medical diagnosis and image classification. The primary goal of classification is to predict a category or class ‘Y’ from some set of inputs X. Among the availability of various classification algorithms, three major methods are depicted here.

A. Decision Tree

Tree based learning algorithms empower predictive models with high accuracy, stability and ease of interpretation; hence they are considered to be one of the best and mostly used supervised learning methods.

A decision tree which uses a tree-like model of decisions can be used to visually and explicitly represent the results and used for decision making procedures and applications. The decision tree model works for both categorical and continuous dependent variables.

1) Key Parameters of Tree Modeling: In decision trees, over-fitting occurs when the tree is designed so as to perfectly fit all samples in the training data set and it can be prevented in two ways *[11].*

Setting constraints on tree size: This can be done by using various parameters which are used to define a tree. The parameters described below are regardless of a tool.

- Minimum samples for a node split
- Minimum samples for a terminal node (leaf)
- Maximum depth of tree (vertical depth)
- Maximum number of terminal nodes
- Maximum features to consider for split

Tree pruning: First the decision tree is created to a large depth. Then we start at the bottom by removing leaves which are giving negative yields when compared from the top.

There are several decision tree algorithms available such as ID3, C4.5, C5.0 & Classification and Regression Trees (CART).

B. K Nearest Neighbor (KNN)

KNN can be used for both classification and regression problems, still, it is more widely used in classification problems. KNN is a simple algorithm that stores all available instances and classifies new instances by a majority of its K neighbors. The new instance being allotted to the class is most common amongst its K nearest neighbors measured by a distance function *[11]*. These distance calculation functions can be Euclidean, Manhattan, Minkowski and Hamming distance. KNN is computationally expensive and variables should be normalized. Otherwise higher range variables can give biased results. Choosing the value of K may be a challenge while accomplishing KNN modeling.

C. Naïve Bayes

It is a classification technique based on Bayes' theorem with an assumption of independence between predictors. A Naive Bayes classifier assumes that the presence of a specific feature in a class is unrelated to the presence of any other feature. Naïve Bayes model is easy to build and it is beneficial for large data sets *[11].* Bayes theorem provides a way of calculating posterior probability and the equation is given as follows.

$P(c|X) = P(X|c)\ P(c) / P(X).$

$P(c|X) = P(x_1|\ c)\ x\ P(x_2|\ c)\ x\ \ldots\ldots x\ P(x_n|\ c)\ x\ P(c)$

Here,

- $P(c|x)$ is the posterior probability of *class* (*target*) given *predictor* (*attribute*).
- $P(c)$ is the prior probability of *class*.
- $P(x|c)$ is the likelihood which is the probability of *predictor* given *class*.
- $P(x)$ is the prior probability of *predictor*.

IV. METHODOLOGY

For the comparative analysis of the classification methods namely decision tree, K nearest neighbour and Naïve Bayes the implementation procedures are discussed here with the datasets used for this study.

A. Tool Used

R is an open source tool to analyse or plot data. It is used to build a statistical model. R has massive number of powerful algorithms implemented as packages. It is most suitable for applied machine learning.

B. Data Sets

The data sets for this work are acquired from UCI Machine Learning repository. Iris data and Breast Cancer Wisconsin (Diagnostic) data sets are considered here for comparative study of three classification methods. Iris data set has 150 instances with four attributes. Here the target is to predict the species class type. Breast Cancer Wisconsin (Diagnostic) data set has 569 instances with 32 attributes. Here the outcome to be predicted is 'M' (Malignant) or 'B' (Benign). In both the datasets 70% of instances have been taken as training data and the remaining 30% has been considered as test data. In R language set.seed() function is used to consider same set of instances all the time. After using set.seed() the instances were selected randomly using sample() function. So, this work shows that the set of instances/rows selected for comparative analysis is same in all the three classification methods for both the datasets.

V. IMPLEMENTATION OF CLASSIFICATION METHODS IN R ENVIRONMENT

The steps to be followed in the implementation of machine learning classification algorithms are given as follows.

- Collection of data
- Pre-processing of data
- Normalise the data (if required)
- Divide the dataset into training and testing dataset
- Train the model
- Plot the model
- Compute the prediction statistics from confusion matrix

By applying the above-mentioned steps in R environment for both the datasets the classification procedure is implemented in this work by using the suitable algorithms.

A. Decision Tree (J48 and C5.0) Algorithms

The decision tree implementation has been carried out by using two algorithms, first algorithm is known as J48() command of RWeka package in R which is an implementation of C4.5 and the second algorithm is C5.0. It is the latest version of the decision tree algorithm and it differs from C4.5 in its code implementation.

1) c5.0() for Iris Dataset: The visualization of decision tree created for Iris Dataset which is implemented using C5.0 algorithm is shown below.

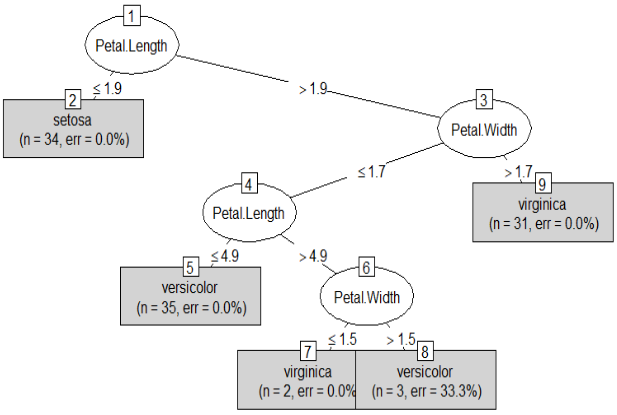

Figure 1. Decision Tree Model Plotting for Iris Dataset

The prediction statistics of this model is found through the confusion matrix for 45 test data.

	setosa	versicolor	virginica
setosa	16	0	0
versicolor	0	12	1
virginica	0	1	15

Prediction Accuracy is calculated as follows.

Accuracy = Number of correct Predictions / Total number of instances for prediction

= 43/45

= 95.56%

Error Rate = Number of wrong predictions / Total number of instances for prediction

= 2/45

= 4.44%

2) c5.0() for Breast Cancer Dataset: For Breast Cancer Dataset the decision tree is created and it is implemented using C5.0 algorithm in R.

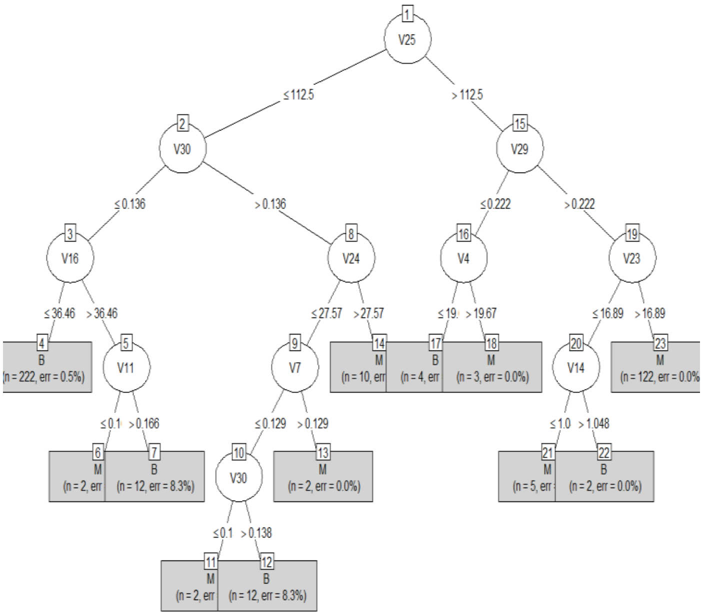

Figure 2. Decision Tree Model Plotting for Breast Cancer Dataset

The prediction statistics of breast cancer data using C5.0 is given as follows.

		Predicted Class	
		B	M
Actual Class	B	103	5
	M	8	55

Prediction Accuracy = 158/171

= 92.4%

B. K Nearest Neighbor (KNN) Algorithm

To implement KNN algorithm the data should be normalised to avoid bias in the results. The variables can be normalised using max-min normalization technique.

$$X' = (X-\min_A) / (\max_A - \min_A)$$

The function in R for normalization procedure is given below.

```
Data_norm<-function(x)
{
 ((x-min(x))/(max(x)-min(x)))
}
```

The normalized datasets are then divided in to training and test data. knn() function in R is used for implementation. The syntax of knn() function is knn(training data tuples, testing data tuples, training targets, k).

1) knn() for Iris Dataset: The confusion matrix framed from knn() method for Iris Dataset is given below.

```
            setosa  versicolor virginica
setosa        16        0        0
versicolor     0       12        2
virginica      0        1       14
```

Here the prediction accuracy is calculated as 42/45 = 93.33%.

2) knn() for Breast Cancer Dataset: The prediction statistics framed through knn() method for Breast Cancer Dataset is given below.

```
      B   M
  B 107   7
  M   1  56
```

The accuracy in prediction is calculated as 163/171 = 95.32%

C. Naïve Bayes Classifier

To implement Naïve Bayes model the dataset is converted as frequency table. Then the likelihood table is created by finding the probabilities. Then the Naïve Bayes equation is used to calculate posterior probability.

1) naïveBayes() for Iris Dataset: The classification accuracy of Naïve Bayes classifier is given with confusion matrix.

```
            setosa  versicolor virginica
setosa        16        0        0
versicolor     0       12        1
virginica      0        1       15
```

Prediction Accuracy = 43/45 = 95.56%

2) naïveBayes() for Breast Cancer Dataset: The confusion matrix framed by using naiveBayes() function on breast cancer dataset is given here with its accuracy calculation.

After implementing in R environment, the output of the classification is given as follows.

```
    B   M
B 103   5
M   7  56
```

Prediction Accuracy = 159/171 = 92.98%

VI. RESULTS AND DISCUSSION

The comprehensive study on three classification methods namely Decision tree, K Nearest neighbour and Naïve Bayes classifiers were carried out in this work. The comparative analysis of these methods is depicted here based on their accuracy in performing classification tasks.

TABLE I. Comparison of Prediction Accuracy of Three Classification Algorithms on Two Different Datasets

S. No.	**Type of Classification Algorithm**	**Prediction Accuracy in Datasets Used for Implementation Purpose**	
		Iris Dataset	**Breast Cancer Dataset**
1	Decision Tree - c5.0()	95.56%	92.4%
2	K Nearest Neighbor - knn()	93.33%	95.32%
3	Naïve Bayes - naiveBayes()	95.56%	92.98%

For the classification task on iris dataset, decision tree and Naïve Bayes methods are showing higher accuracy than KNN. But for breast cancer dataset KNN algorithm shows highest accuracy. So, on an average K Nearest Neighbor algorithm shows higher accuracy when both the datasets are considered together. These two different datasets selected here for performance comparison are having different types of attributes and number of instances also not similar. The accuracy of the outcome will differ based on the number of instances and attributes (features) to be considered for classification tasks. So, the calculated prediction accuracies of these three classification algorithms may differ when they are implemented on other data sets. This work can be enhanced on other different types of inputs but the type of attributes, number of attributes and instances of datasets used may be similar. So that the comparative study on classification algorithms will be more efficient and effective.

VII CONCLUSION

The employment of three classification algorithms decision tree, K nearest neighbour and Naïve Bayes classifiers on two different datasets of Iris and Breast Cancer show that, on an average K Nearest Neighbour procedure is having more accuracy than other two methods. But it is evidently based on the type of datasets to be used for implementation purpose. The accuracy of classification depends on the type of variables used and also it is based on the applications and requirements. The results which are found using R environment may differ while we use some other tools for implementation. Hence based on the constraints, requirements and type of input datasets specific algorithm and tool can be chosen. The comparative analysis based on the performance metrics can be further enhanced on some other real-time datasets in future.

VII. REFERENCES

[1]. B Nithya, "An Analysis on Applications of Machine Learning Tools, Techniques and Practices in Health Care System", International Journal of Advanced Research in Computer Science and Software Engineering 6(6), June-2016, pp. 1-8.

[2]. Pellakuri et al., "Performance Analysis and Optimization of Supervised Learning Techniques for Medical Diagnosis Using Open Source Tools", International Journal of Computer Science and Information Technologies, Vol. 6 (1), 2015, 380-383.

[3]. Sujata, Priyanka, "Performance Analysis of Different Classification Methods in Data Mining for Diabetes Dataset Using WEKA Tool", International Journal on Recent and Innovation Trends in Computing and Communication, Volume: 3 Issue: 3, Mar' 2015.

[4]. Subrata Kumar, "Performance Analysis of Data Mining Algorithms For Breast Cancer Cell Detection Using Naïve Bayes, Logistic Regression and Decision Tree", International Journal of Engineering And Computer Science ISSN: 2319-7242, Volume 6, Issue 2, Feb. 2017, Page No. 20388-20391.

[5]. Sayali D. Jadhav, H. P. Channe, "Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques", International Journal of Science and Research, Volume 5 Issue 1, January 2016.

[6]. Bhrigu Kapur et al., "Comparative Study on Marks Prediction using Data Mining and Classification Algorithms", International Journal of Advanced Research in Computer Science, Volume 8, No. 3, March – April 2017.

[7]. Sara Silva et al., "A Comparison of Machine Learning Methods for the Prediction of Breast Cancer", EvoBIO 2011, LNCS 6623, pp. 159–170, Springer-Verlag Berlin Heidelberg 2011.

[8]. Vanneschi et al., "A comparison of machine learning techniques for survival prediction in breast cancer", BioData Mining 2011, 4:12.

[9]. Leonardo et al., "Identification of Individualized Feature Combinations for Survival Prediction in Breast Cancer: A Comparison of Machine Learning Techniques", EvoBIO 2010, LNCS 6023, pp. 110–121, Springer-Verlag Berlin Heidelberg 2010.

[10]. Liyang Wei et. al., "A Study on Several Machine-Learning Methods for Classification of Malignant and Benign Clustered Microcalcifications", IEEE Transactions on Medical Imaging, Vol. 24, No. 3, March 2005.

[11]. https://www.analyticsvidhya.com

[12]. Brett Lantz, "Machine Learning with R", 2nd Edition, PACKT Publishing.