Scientific Journal of Impact Factor (SJIF): 4.14
e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 3, Issue 2, February -2016

Comparative Study of Machine Learning Algorithms for Classification of Datasets using R Programming

Ramaswamy M[1], Savitha B[2]

[1] Assistant Professor, Department of IT, Rajalakshmi Institute of Technology, Chennai,
[2] Assistant Professor, Department of IT, Rajalakshmi Institute of Technology, Chennai.

Abstract — *Over the past years, periodically many organizations have started to capture large volumes of historical data for describing their operations, products, and customers. Data Mining apparently tries to extract knowledge or some unknown interesting patterns from these huge unstructured data. During this process machine learning algorithms are used. The aim of this paper is to study various machine learning algorithms for classification and to compare them. In this paper C5.0, SVM, Random Forest, GBM, Bayes Classifier, MARS, AdaBoost and Deep Learning have been compared by using the various publically available datasets. The R Programming language has been used for experimenting all the algorithms.*

Keywords: GBM, MARS, Deep Learning

I. INTRODUCTION

Machine Learning is an application of artificial intelligence where available information is used to process/assist the manipulation of statistical data by providing a collection of data analysis technique. It involves concepts of automation and a high level of generalization in order to get a system that performs well on yet unseen data instances. It has some of the very well established statistical methods like logistic regression and principal component analysis, while many others are not. It provides a more broad class of flexible alternative analysis methods which suit much better for modern sources of data. Supervised and Unsupervised machine learning are two main classes of machine learning techniques. Mitchell [1] describes a broad range of machine learning algorithms which are based on the statistical principles. Classification and Prediction are two forms of data analysis that can be used to extract patterns describing important data classes or to predict future data trends [2]. Classification is predicting a certain outcome based on a given input. In order to predict, the algorithm processes a training set containing a set of attributes and the respective outcome is the prediction attribute. Then by discovering the relationships between the attributes make it possible to predict the outcome. Next the algorithm is given a data set not seen before, called prediction set, which contains the same set of attributes, except for the prediction attribute which is not yet known. The algorithm analyses the input and produces a prediction. The effectiveness of the algorithm depends on the prediction accuracy.

II. RELATED WORKS

For the extraction of knowledge from the database, a large number of algorithms have been developed in recent years, especially for classification related jobs. Yao [3] et al. applied a modified version of C4.5 called R-C4.5, which improved the efficiency of attribution selection and model partitioning. Jaya [4] et al discussed how the data classification is based on supervised machine learning algorithms which result in accuracy, time taken to build the algorithm. Colombet [5] et al implemented and evaluated the performance of CART and artificial neural networks to predict the risk of cardiovascular disease in a real database. A better performance of Bayes algorithm has been showed by Shadmehr et al [6]. Sonu et al [7] proposed a neural network model for prediction of diabetes based on 13 early symptoms of the disease. Breiman [8] presented the idea of Random Forests which perform well as compared with other classifiers and overcomes the over fitting problem. Santi Waulan et al [9] proposed a new SSVM for classification problems, which effectively classified diabetes disease. Stephen Tyree [10] et al has used Gradient Boost Regression Trees for ranking the web searches. Better performance of HONN (Higher Order Neural Network) has been found by Spikvoska et al [11]. Provost et al [12] examined the issue of predicting probabilities of decision trees.

III. METHODOLOGY

3.1 C5.0 Algorithm

C5.0 [24] is a commercial version of C4.5 which is widely used. It is generally used for large datasets. The decision tree induction is close to that C4.5, but the rule generation is different. It is also better in terms of efficiency when compared with C4.5. Here, based on the fields the samples are splitted, which provides maximum information gain and the splitting continues until the sample subset cannot be split and is usually according to another field. Finally, examine the lowest level split, those sample subsets that don't have remarkable contribution to the model will be rejected.

3.2 Random Forest Algorithm

Random Forest is an ensemble supervised machine learning technique, which has tremendous potential because its performance which is found to be comparable with ensemble techniques bagging and boosting [18]. In a random forest, each node is split using the best among a subset of predictors randomly chosen at that node. SVM performs very well when compared to many other classifiers, including discriminant analysis, support vector machines and neural networks, and is robust against over fitting [19].

3.3 Support Vector Machines

The Support Vector Machine was first proposed by Vapnik [16]. Recent works have reported that the classification capability of SVM's are far more than many other data classification techniques. Regression and classification tasks are performed by constructing nonlinear decision boundaries. Wide range of classification and regression tasks can be handled by SVM due to its large degree of flexibility. There are several types of Support Vector models including linear, polynomial, RBF, and sigmoid.

3.4 Gradient Boosting Machines

Gradient boosting machines are a family of powerful machine-learning techniques which has a wide range of practical applications. They can be highly customized to suit various particular application needs and can perform regression, classification and ranking. The idea of gradient boosting originated in the observation by Leo Breiman [13] that boosting can be interpreted as an optimization algorithm on a suitable cost function. Explicit regression gradient boosting algorithms were subsequently developed by Jerome H. Friedman [14] [15] simultaneously with the more general functional gradient boosting perspective of Llew Mason.

3.5 Naive-Bayes Classifier

A Bayesian Classifier [17] is a probabilistic model in which the classification is a latent variable. Here the variable is probabilistically related to the observed variable. The simplest case is the naive Bayesian classifier, which makes the independence assumption that the input features are conditionally independent of each other given the classification. The independence of the naive Bayesian classifier is embodied in a particular belief network where the features are the nodes, the target variable (the classification) has no parents, and the classification is the only parent of each input feature. This belief network requires the probability distributions P(Y) for the target feature Y and P(Xi|Y) for each input feature Xi. For each example, the prediction can be computed by conditioning on observed values for the input features and by querying the classification.

3.6 Multivariate Adaptive Regression Splines (MARS)

Multivariate Adaptive Regression Splines (MARS) [23] is flexible regression modeling of high dimensional data. It takes the form of an expansion in product spline basis functions, where the number of basis functions as well as the parameters associated with each one are automatically determined by the data. This procedure is motivated by the recursive partitioning approach to regression and shares its attractive properties. Unlike recursive partitioning, however, this method produces continuous models with continuous derivatives. It has more power and flexibility to model relationships that are nearly additive or involve interactions in at most a few variables. In addition, the model can be represented in a form that separately identifies the additive contributions and those associated with the different multivariable interactions.

3.7 AdaBoost Algorithm

The AdaBoost algorithm of Freund and Schapire [22] was the first practical boosting algorithm, and remains one of the most widely used and studied, with applications in numerous fields. AdaBoost is adaptive in the sense that subsequent weak learners are tweaked in favor of those instances misclassified by previous classifiers. It is sensitive to noisy data and outliers. But less susceptible to the over fitting problem that occur in other learning algorithms. The individual learners can be weak, but as long as the performance of each one is slightly better than random guessing, the final model can be proven to converge to a strong learner.

3.8 Deep Learning Technique

Deep Learning Technique [20] is a new area of Machine Learning research, which has been introduced with the objective of moving Machine Learning closer to one of its original goals: Artificial Intelligence. Unlike the neural networks of the past, modern Deep Learning has cracked the code for training stability and generalization and scales on big data. It is often the algorithm of choice for highest predictive accuracy, as deep learning algorithms performs quite well in a number of diverse problems. One of the promises of deep learning is replacing handcrafted features with efficient algorithms for unsupervised or semi-supervised feature learning and hierarchical feature extraction [21].

IV. EXPERIMENTAL RESULTS

The data used here for the comparison are publically available in the UCI Machine Learning Repository [15]. Table 1 shows the characteristics of the datasets used. Prior to the evaluation of the performance of the algorithms, the datasets are preprocessed since some of the algorithms cannot handle continuous values or missing values. The R-3.2.3 version and CRAN Packages like C50, randomForest, e1071, gbm, earth, adabag, h2o.ai have been used for this comparative study purpose. All the datasets used for this study have been divided into test data and train data based on 70:30 ratio. The classification accuracy results of the above algorithms over the datasets are show in Table 2. The comparative chart of the classifier algorithms is show in Fig.1.

S.No.	Datasets	No. of Attributes	Data Set Characteristics	No. of Instances
1	Diabetic Retinopathy Debrecen Data Set	20	Multivariate	1151
2	Pima Indians Diabetes Data Set	8	Multivariate	768
3	Chess Data Set	36	Multivariate	3196
4	Car Evaluation Data Set	6	Multivariate	1728
5	Liver Disorders Data Sets	7	Multivariate	345

Table-1 Dataset Characteristics

S.No.	Algorithms	Dataset-1	Dataset-2	Dataset-3	Dataset-4	Dataset-5
1	C5.0 Algorithm	85.49	79.88	94.13	62.28	91.86
2	Random Forest Algorithm	92.18	89.65	80.38	83.72	93.06
3	Support Vector Machines	73.47	72.54	81.5	64.21	79.94
4	Gradient Boosting Machines	81.13	78.5	81.23	82.85	71.51
5	Naive-Bayes Classifier	62.91	67.36	75.23	82.47	56.69
6	MARS	75.5	69.55	79.66	61.07	72.09
7	AdaBoost algorithm	71.52	82.75	87.22	87.5	89.24
8	Deep Learning	73.59	83.05	83.5	96.7	78.85

Table-2 Dataset Characteristics

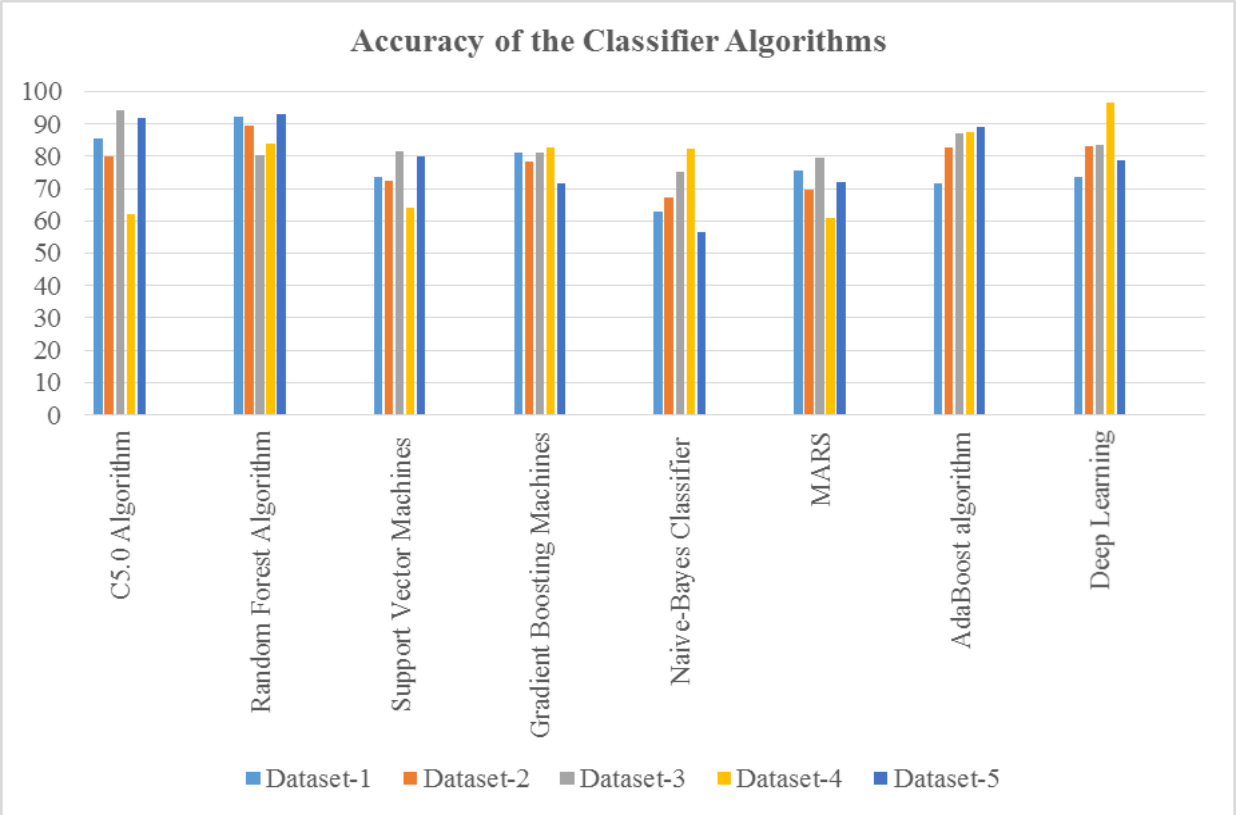

Fig. 1 Comparison Chart

V. CONCLUSION

Classification Techniques have always been an interesting area for many researchers. It is also an important area in Data Mining. In recent days, Machine Learning has also gaining much more popularity. From the comparative study, Random Forest, GBM and Deep Learning Algorithm yield a better classification rate and they seem to have a consistent performance over all the experimental data used.

REFERENCES

[1] Mitchell, T. Machine Learning. McGraw-Hill, New York, 1997.

[2] H. Jiawei and K. Micheline, (2008) Data Mining-Concepts and Techniques, Second Edition, Morgan Kaufmann - Elsevier Publishers, ISBN: 978-1-55860-901-3.

[3] Yao, Z.; Lei, L.; Yin, J., "R-C4.5 Decision tree model and its applications to health care dataset". Proceedings of International Conference on Services Systems and Services Management 2005, pp. 1099-1103.

[4] Jaya Rama Krishnaiah.V.V., K.Ramchand H Rao, "Predicting the Heart attack symptoms using Biomedical data mining techniques", International Journal of Computer Science & Applications, Volume 1, No. 3, May 2012.

[5] Colombet, I., Ruelland, A., "Models to predict cardiovascular risk: comparison of CART, multilayer perceptron and logistic regression". Proceedings of AMIA Symp 2000, p 156-160.

[6] R. Shadmehr and Z. D'Argenio, "A comparison of a neural network based estimator and two statistical estimators in a sparse and noisy environment", in IJCNN-90 Proceedings of the international joint conference on neural networks, 289-292, IEEE Neural Networks Council.

[7] Sonu Kumari and Archana Singh, "A Data Mining Approach for the Diagnosis of Diabetes Mellitus", Proceedings of 7th International Conference on Intelligent Systems and Control, 2013.

[8] Breiman, L., Random Forests, Machine Learning 45(1), 5-32, 2001.

[9] S. W. Purnami and S. P. Rahayu, "A New Smooth Support Vector Machine and Its Applications in Diabetes Disease Diagnosis," Journal of Com-puter Science, Vol. 5, No. 12, pp. 1006-1011.

[10] Stephen Tyree et al, "Parallel Boosted Regression Trees for Web Search Ranking" in ACM WWW 2011, March 28–April 1, 2011.

[11] L. Spikovska and M.B. Reid, "An empirical comparison of ID3 and HONNS for distortion invariant object recognition" in Proceedings of the 2nd International IEEE conference, Los Alamitos, CA, IEEE Computer Society Press, 1990.

[12] F. Provost et al , "Efficient progressive sampling", Fifth ACM SIGKDD, International conference on knowledge Discovery and Data Mining, San Diego, USA, 1999.

[13] https://www.stat.berkeley.edu/~breiman/arcing-the-edge.pdf

[14] Friedman, J. H.," Greedy Function Approximation : A Gradient Boosting Machine " in Annals of Statistics 2001, Vol. 29, No. 5, 1189–1232

[15] https://statweb.stanford.edu/~jhf/ftp/stobst.pdf

[16] V. Vapnik. The Nature of Statistical Learning Theory. NY: Springer-Verlag. 1995.

[17] http://artint.info/html/ArtInt_181.html

[18] R. Shapire, Y. Freund, P. Bartlett, and W. Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. Annals of Statistics, 26 (5):1651–1686, 1998.

[19] L. Breiman. Bagging predictors. Machine Learning, 24 (2):123–140, 1996.

[20] Deep Learning Tutorial(Release 0.1), LISA lab, University of Montreal.

[21] Lee, S. Y. (2013). "Hierarchical Representation Using NMF". Neural Information Processing. Lectures Notes in Computer Sciences 8226. Springer Berlin Heidelberg. pp. 466.

[22] Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences 55(1), 119–139 (1997).

[23] "Multivariate Adaptive Regression Splines" by Jerome H. Friedman in SLAC PUB-4960 Rev, Tech Report 102 Rev, August 1990

[24] Jiawei Han , Micheline Kamber " Data Mining – Concepts and Techniques" Elsevier, 2007 pages 291- 310.

