An Empirical Study on Hybrid Classification Algorithms

Edagotti Pavithra[1], Anjan Babu G[2]

[1]PG Student, Department of Computer Science, Sri Venkateswara University, Tirupati
[2]Professor, Department of Computer Science, Sri Venkateswara University, Tirupati

Abstract— *Information mining is that the technique for breaking down information from totally different views and summing up it into helpful data. Characterization could be a data handling method upheld AI which is utilized to classify everything in a bunch of information into a gathering of predefined classes or teams. Order is technique for summing up the information predictable according to various occurrences. Order calculations as a significant innovation in information mining and AI have been generally examined and applied. Numerous strategies can be utilized to fabricate classifiers, for example, the choice tree, Bayesian strategy, case-based learning, counterfeit brain organization and backing vector machine. This paper centers around the grouping techniques in view of NB Tree learning and BF Tree, Dermatology informational collection was utilized for the arrangement with 366 cases with 65 qualities as autonomous variable and one as reliant variable for the examination. The outcomes show that NB Tree (95.628%) viewed as the calculation with most accuracy and precision when contrasted with BF Tree (93.98%) calculation.*

I. INTRODUCTION

Data mining is a technology that offers extracting or discovering new relations, hidden knowledge and important patterns from such data. It is also known as Knowledge Discovery in Databases (KDD). Data mining technique is important for analysis purpose. Data mining supports different techniques such as classification, clustering, association rule mining, outlier analysis etc [1][4]. Data Mining (DM) discovers hidden relationships in data, in fact it is a part of wider process called "knowledge discovery". Knowledge discovery describes the phases which must be done to ensure reaching meaningful results through research. The objective of DM process is to obtain information out of a dataset and converts it into a comprehensible outline. An understanding of algorithms is combined with detailed knowledge of the dataset an understanding of algorithms is combined with detailed knowledge of the datasets. Data mining must afford very complex and different situations to reach quality solutions. Therefore, data mining is a research field where many advances are being done to accommodate and solves merging problems [1]. For present study purpose classification technique is investigated.

II. CLASSIFICATION

Classification plays an important role in data mining and machine learning. The purpose of classification algorithm is to construct a classifier, and then analyzes the characteristics of the unknown data to get an accurate model. The performance of the classifier is measured by its classification accuracy. Building effective classification systems is one of the central tasks of data mining. The main purpose of supervised learning is to build a simple and unambiguous model of the allocation of class labels in terms of predictor features [2][7]. The classifiers are then used to classify class labels of the testing instances where the values of the predictor features are known, to the value of the class label which is unknown [3][5]. Classification of this tremendous amount of data is time consuming and utilizes excessive computational effort, which may not be appropriate for many applications.

III. METHODOLOGY

Many different types of classification techniques have been proposed in literature that includes Decision Trees, Naïve Bayesian methods, Neural Networks, Logistic Regression, SVM and KNN etc. In this paper, we evaluate the performance of the NB Tree algorithms on Dermatology data set was used for the classification compared with the BF Tree algorithm.

3.1 BF Tree Algorithm

In best-first top-down induction of decision trees, the best split is added in each step (e.g. the split that maximally reduces the Gini index). This is in contrast to the standard depth-first traversal of a tree. The resulting tree will be the same, just how it is built is different. The objective of this project is to investigate whether it is possible to determine an appropriate tree size on practical datasets by combining best-first decision tree growth with cross-validation-based selection of the number of expansions that are performed. Pre-pruning, post-pruning, CART-pruning can be performed this way to compare.

3.2 NB Tree

BTree and NBTree is a hybrid algorithm that represents a cross between Naive Bayes classifier and C4.5 Decision Tree classification and it's best described as a decision tree with nodes and branches [9]. The NBTree algorithm is written below with input of T sets of labeled instances and a decision-tree with Naive Bayes category at the output (leaves):

1. For each attribute Xi, evaluate the utility, u(Xi), of a split on attribute Xi. For continuous attributes, a threshold is also evaluated at this stage.
2. Let J = AttMax(Ui). The attribute with highest utility (Maximum utility).
3. If Uj is not significantly better than the utility of the current node, create a Naive Bayes classifier for the current node and return.
4. Partition T according to the test on Xj. If Xj is continuous, a threshold split is used; if Xj is discrete, a multi-way split is made for all possible values.
5. For each child, call the algorithm recursively on the portion of T that matches the test leading to the child

IV. EXPERIMENTAL RESULTS

The analyses have been directed by utilizing R programming Language. R is a sophisticated statistical software package, which provides new approaches to data mining., it is an open-source tool for analysis of data mining algorithms. The R Language is a bundle for information characterization, grouping and representation. We have considered the Pima diabetes from the UCI Machine Learning Repository data sets for assessing the productivity and adequacy of decision tree calculation [8]. The characteristic data information is consolidated in Table-1. The standard dataset is parceled into two sets one for training (75%) and another set for testing (25%).

TABLE 1
DATASET INFORMATION

S. No	Name of the Dataset	No. of Attributes	No. of Instances	No. of Classes
1	Dermatology	65	366	6

We survey our two models using assorted execution estimations like Accuracy, Precision and Recall, the Experimental results are showed up in the table-1 and same showed up in the Figure-1.

TABLE 2
PERFORMANCE OF CLASSIFIERS

Algorithm	Accuracy	Precision	Recall
BF Tree	93.989	94.1	94
NB Tree	95.628	95.8	95.6

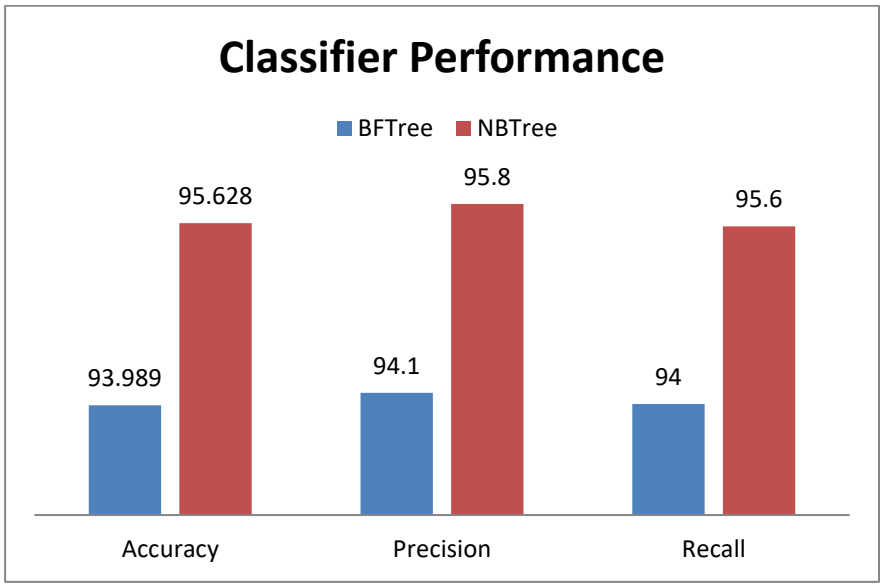

Figure-1: Experimental Results

We find in the Figure-1, the introduction of the NB Tree estimation has accomplished 95.628% precision and BF Tree has achieved 93.989%, As the result from assessment among the two computations, we find that most vital precision of Classification model is NB Tree (95.628%). So, the NB Tree algorithm have got highest accuracy, with a 1.639% difference when compared to BF Tree algorithm.

V. CONCLUSION

The clinical dataset in the various information mining and the artificial intelligence techniques are open and from there on the gigantic piece of clinical information mining is to develop the accuracy and suitability of disease finding. In this paper, three datamining strategy learning calculation for dermatology jumble figure has been framed. The evaluation the sensibility of the methodology utilizing undeniable arrangement metric appraisal has been made and it has been shown that the precision of the model was moved along. To see dermatology disease from monstrous dataset, affirmation assessment pointlessly more proficient. In this manner NB Tree (95.628%) classifier is proposed for examination of clinical confirmation presumption based solicitation to additionally foster outcomes with precision and execution.

REFERENCES

[1] D. Hand, H. Mannila, P. Smyth.: Principles of Data Mining. The MIT Press. (2001)

[2] G. Ravi Kumar and K. Nagamani, “Banknote Authentication System utilizing Deep Neural Network with PCA and LDA Machine Learning Techniques”, International Journal of Recent Scientific Research, ISSN: 0976-3031, Volume 9, Issue 12(D), PP:30036-30038, 2018

[3] Ian H. Witten and Eibe Frank. Data Mining: Practical machine learning tools and techniques.2nd ed. San Francisco: Morgan Kaufmann, 2005.

[4] J. Han and M. Kamber,” Data Mining concepts and Techniques”, the Morgan Kaufmann series in Data Management Systems, 2nd ed. San Mateo, CA; Morgan Kaufmann, 2006.

[5] N. Michael, “Artificial Intelligence - A Guide to Intelligent Systems”,2nd edition, Addison Wesley, 2005.

[6] P.-N. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining. Reading, MA: Addison-Wesley, 2005.

[7] S.Rahamat Basha and G.Ravi Kumar Surya Bhupal RaoG,”A Comparative approach of Text Mining: Classification, Clustering and Extraction Techniques”, Journal of Mechanics of Continua and Mathematical Sciences, ISSN: 2454 -7190, Special Issue, No.-5,PP:120-131, 2020

[8] UCI machine learning repository. http://archive.ics.uci. edu/ml/

[9] Pumpuang P., Srivihok A. , Praneetpolgrang P. , "Comparisons of Classifier Algorithms: Bayesian Network, C4.5, Decision Forest and NBTree for Course Registration Planning Model of Undergraduate Students", SMC 2008. IEEE International Conference.