An Implementation of Efficient Datamining Classification Algorithm using Nbtree

A.Veeraswamy
Research Scholar
VELTECH Dr.RR & Dr.SR
TECHNICAL UNIVERSITY
Chennai, Tamil Nadu, India.

S.Appavu alias Balamurugan, PhD.
Professor & Research Coordinator
Department of ECE
KLN College of Information Technology, Madurai, Tamil Nadu, India.

E.Kannan, PhD.
Registrar & Dean (Academics)
VELTECH Dr.RR & Dr.SR
TECHNICAL UNIVERSITY
Chennai, Tamil Nadu, India.

ABSTRACT

Knowledge [no more Information] is not only power, but also has significant competitive advantage. Data warehousing is not a new idea. The use of corporate data for strategic decision making, as opposed to the use of data for tracking and enabling operations, has gone on for a computing itself. As the business these days contain huge amounts of data and the users connected to these databases across the globe and round the clock have the necessity for maintaining a separate database for the sake of analysis. This paper proposes one method of feature selection of NB Tree Algorithm. The Proposed algorithm (NB Tree) gives an effective Classification Algorithm for reducing computational time and gives better accuracy results compare with another algorithms. In many applications, however, an accurate ranking of instances based on the class probability is more desirable. The dependence between two attributes is determined based on the probabilities of their joint values that contribute to true and false classification decisions. The paper also evaluates the approach by comparing it with existing feature selection algorithms over 8 datasets from University of California, Irvine (UCI) machine learning databases. The proposed method shows better results in terms of number of selected features, classification accuracy, and running time than most existing algorithms.

Keywords

Feature selection, Data mining, Classification, J48, Decision Tree, and NB Tree.

1. INTRODUCTION

Classification is one of the fundamental problems in data mining. In classification, the goal is to learn a classifier from a given set of instances with class labels, which correctly assigns a class label to a test instance. The performance of a classifier is usually measured by its classification accuracy. Classification has been extensively studied and various learning algorithms have been developed, such as decision trees and Bayesian networks that can be categorized into two major approaches: probability-based approaches and decision boundary-based approaches [1].

Feature selection is one of the important and frequently used techniques in data preprocessing for data mining. It reduces the number of features, removes irrelevant, redundant, or noisy data, and brings the immediate effects for applications, speeding up a data mining algorithm, improving mining performance such as predictive accuracy and result comprehensibility. Feature selection is a fertile field of research and development in statistical pattern recognition, machine learning, and data mining since the 1970s, and widely applied to many fields such as text categorization, image retrieval, customer relationship management, intrusion detection, and genomic analysis. Feature selection is a process that selects a subset of original features. The optimality of a feature subset is measured by an evaluation criterion [2].

In this paper, we give an overview of the popularly used feature selection algorithms under a united framework. Moreover, we propose a novel classification algorithm based on the all the features for determining the dependent attributes in a dataset and removing those dependent attributes, thereby reducing the attribute set to increase the classification accuracy and reduce the computational time. Experiments on real world datasets show that the proposed method is favorable in terms of its Accuracy and Efficiency when compared with other state-of-art algorithms.

Data pre-processing is an often neglected but important step in the data mining process [4]. The phrase "Garbage In, Garbage Out" is particularly applicable to data mining and machine learning projects. Data gathering methods are often loosely controlled, resulting in out-of-range values, impossible data combinations, missing values, etc. The data preprocessing having the algorithms of Data Cleaning ,Data Transformation, Data Reduction, Data Integration and Normalization [3]. The Data Cleaning is the process of removing noisy data, inconsistent and redundant data. In Data transformation the data will be converted in to the data mining process. The Data integration is the process of combining the data bases in to the data ware house. After the preprocessing, the data will be given to the data mining process. In this paper introduces the attribute subset selection algorithm for data preprocessing.

2. RELATED WORK

Feature selection is a mature area of research. We will present a brief overview of the different feature selection methods. Blum and Langley [2] classified the feature selection techniques into three basic approaches. In the firsrst approach, known as the embedded approach, a basic induction method is used to add or remove features from the concept description in

response to prediction errors on new instances. The second approach is known as the filtering approach, in which, various subsets of features are explored to find an optimal subset, which preserves the classification. The third approach is known as wrapper methods which evaluate alternative feature sets by running some induction algorithm on the training data and using the estimated accuracy of the resulting classifier as its metric [2].

3. PROPOSED WORK

NB Tree decision tree is a cross between Naive Bayes classifier and classification. NB Tree model Best described as a decision tree with nodes and branches Leaf nodes of the Bayesian classification. As with other tree-based Classification, NB Tree through, with branches and nodes. Given the A set of instances of the algorithm evaluation node "Practice" for each division of the property. If the greatest value the property is significantly better than the practices Instance, based on the current node, will be divided into Property. If you do not divide, providing an important to better the effectiveness of a naive Bayesian classifier to create the current node. The effectiveness of compute nodes discrete data and ten times the cross-validation performed Using Bayesian estimation accuracy [1].

4. SYSTEM IMPLEMENTATION

The proposed algorithm is implemented using Java. The stepwise approach is as follows. The input to the system is given as an attribute-relation file format (ARFF) file. A table is created in Oracle using the name specified in "@relation". The attributes specified under "@attribute" and instances specified under "@data" are retrieved from the ARFF file and then they are added to the created table. This procedure is followed for providing the training set as well as test set. The created table acts as the dataset and is given as the input to the proposed algorithm. The number of predictor attributes and its distinct values and number of distinct values in class attribute are calculated, and these values are used for the calculation of probabilities. The combination of attribute value should occur at least once in the dataset, because while finding the dependency between attribute values if a combination of attribute value did not occur once, then it will lead to alternate zeros resulting in zero probability and dependency that cannot be found. Thus, the above condition is checked before a combination of attribute value is given to the proposed method. The probabilities are calculated for the given input. Based on the probabilities, the dependent attributes are identified.

5. EXPERIMENTAL RESULTS & DISCUSSION

All together 13 datasets are selected from the UCI machine learning repository and the UCI knowledge discovery in databases (KDD) archive A summary of datasets is presented in Table 1. For each dataset, we run all Classification Algorithms Decision Tree, NBTree, J48, IBK, Naive Bayes, on the original dataset as well as each newly obtained dataset containing only the selected features from each algorithm and recorded the overall accuracy by 10 fold cross validation.

Table1: Details description of dataset used in the experiment

S.NO	Name of the Dataset	No. of Instances	No. Of attributes
1	CMC	1473	10
2	HEPATITIS	155	20
3	IONOSPHERE	351	35
4	LABOR	57	17
5	LUNG-CANCER	32	57
6	MUSHROOM	8124	23
7	PIMA_DIABETES	768	9
8	SPONGE	76	46
9	SPAMBASE	4601	58
10	VEHICLE	846	19
11	WAVEFORM	5000	41
12	ZOO	101	18
13	NURSERY	12960	9

Fig. 1 Number of features selected by the proposed method

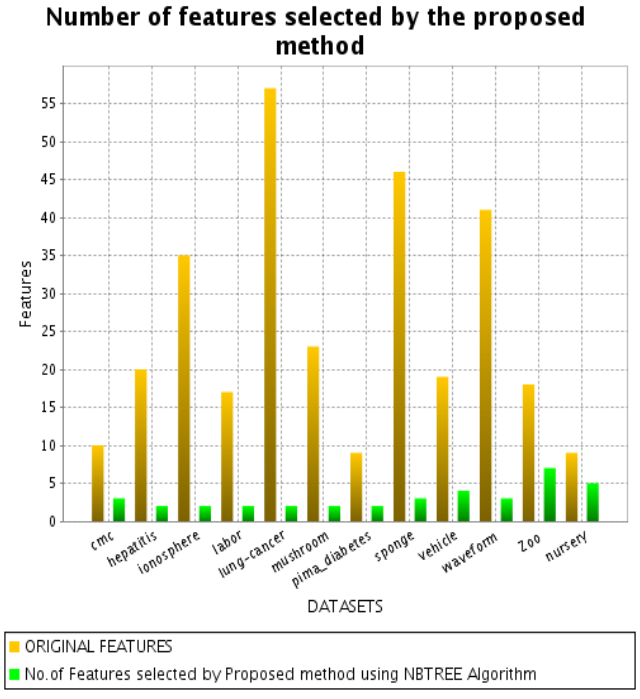
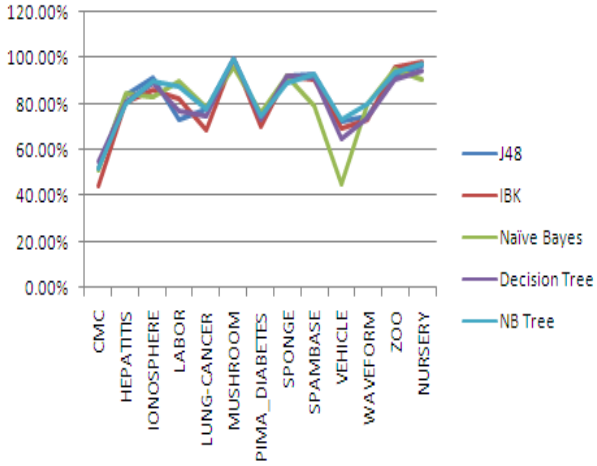


Fig 2: Effect of the proposed Classification Algorithm Comparison of remaining algorithms performance graph



6. CONCLUSION

This paper proposes a NB Tree algorithm can remove redundancy from the original dataset. The main idea provided is to find the dependent attributes and remove the redundant ones among them. We compared all the Classification like J48, IBK, Naïve Bayes, Decision Tree, NB Tree, in this average accuracy of NB Tree Classification algorithm gives better performance comparing to other algorithms see Table 3 and Table 4. We compared the performance of a number of algorithms on the UCI machine learning repository datasets.

7. REFERENCES

[1] R. Kohavi, "Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid," ser. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. AAAI Press, 1996, pp. 202–207.

[2] A. L. Blum, P. Langley. Selection of Relevant Features and Examples in Machine Learning. Artificial Intelligence, vol.97, no. 1 PP 245-271, 1977.

[3] Varun Kumar, Nisha Rathee,"Knowledge Discovery from Database using an Integration of clustering and Classification", IJACSA, vol 2 No.3,PP. 29-33,March 2011.

[4] G.Karraz,G.Magenes,"Automatic Classification of Heart beats using Neural Network Classifier based on a Bayesian Frame Work", IEEE, Vol 1,2006.

[5] R. Kohavi, G. H. John. Wrappers for Feature Subset Selection. Artificial Intelligence, vol. 97, no. 1&2, pp. 273-324, 1997.

[6] Subramanian Appavu Alias Balamurugan, Ramasamy Rajaram.Effective and Efficient Feature Selection for Large-scale Data Using Bayes Theorem, International Journal of Automation and Computing February 2009, 62-71.

Table 3: Accuracy of NB Tree on selected features for each Classification Algorithm

S.NO	Name of the Dataset	J48	IBK	Naïve Bayes	Decision Tree	NB Tree
1	CMC	52.14%	44.33%	50.78%	54.99%	51.73%
2	HEPATITIS	83.87%	80.65%	84.52%	81.29%	80%
3	IONOSPHERE	91.45%	86.32%	82.62%	89.46%	89.74%
4	LABOR	73.68%	82.46%	89.47%	77.19%	87.72%
5	LUNG-CANCER	78.13%	68.75%	78.13%	75%	78.13%
6	MUSHROOM	100%	100%	95.83%	100%	100%
7	PIMA_DIABETES	73.83%	70.18%	76.30%	73.31%	74.35%
8	SPONGE	92.11%	92.11%	92.11%	92.11%	89.47%
9	SPAMBASE	92.98%	90.78%	79.29%	91.41%	93.05%
10	VEHICLE	72.46%	69.86%	44.80%	65.01%	72.93%
11	WAVEFORM	75.08%	73.62%	80%	73.80%	79.88%
12	ZOO	92.08%	96.04%	95.05%	91.09%	94.06%
13	NURSERY	97.05%	98.38%	90.32%	94.69%	97.49%
AVERAGE ACCURACY		**82.68%**	**81.03%**	**79.93%**	**81.48%**	**83.73%**

Table 4: Accuracy of NB Tree Compare with Decision Tree Algorithm on selected features for each Classification Algorithm

S.NO	Dataset	Decision Tree Accuracy	NBTREE Accuracy
1	CMC	54.99%	51.73%
2	HEPATITIS	81.29%	80%
3	IONOSPHERE	89.46%	89.74%
4	LABOR	77.19%	87.72%
5	LUNG-CANCER	75%	78.13%
6	MUSHROOM	100%	100%
7	PIMA_DIABETES	73.31%	74.35%
8	SPONGE	92.11%	89.47%
9	SPAMBASE	91.41%	93.05%
10	VEHICLE	65.01%	72.93%
11	WAVEFORM	73.80%	79.88%
12	ZOO	91.09%	94.06%
13	NURSERY	94.69%	97.49%
Average Accuracy		**81.48%**	**83.73%**