International Journal for Modern Trends in Science and Technology
Volume: 03, Special Issue No: 01, February 2017
ISSN: 2455-3778
http://www.ijmtst.com

Feature Ranking Based Data Classification for Large Scale Data Analysis Using J48 Algorithm

Dr.A.Veeraswamy[1] | A V S Sudhakara Rao[2] | T.Krishna Kishore[3]

[1,2,3]Department of Computer Science & Engineering, St.Ann's College of Engineering & Technology, Chirala, Andhra Pradesh, India.

To Cite this Article
Dr.A.Veeraswamy, A V S Sudhakara Rao and T.Krishna Kishore, "Feature Ranking Based Data Classification for Large Scale Data Analysis Using J48 Algorithm", *International Journal for Modern Trends in Science and Technology*, Vol. 03, Special Issue 01, 2017, pp. 104-107.

ABSTRACT

Data Mining is a technique used in various domains to give meaning to the available data and different types of Data to be handled like numerical data, non-numeric data, image data...etc. In classification tree modelling the data is classified to make predictions about new data. Using old data to predict new data has the danger of being too fitted on the old data. In this we evaluated different types of data to be collected from UCI repository for classify the data using the different classification algorithms J48, Naive Bayes, Decision Tree, IBK. This paper evaluates the classification accuracy before applying the feature selection algorithms and comparing the classification accuracy after applying the feature selection with learning algorithms.

KEYWORDS: *J48, Naïve Bayes, Decision Tree, IBK, Data mining*

I. INTRODUCTION

As computer and database technologies develop rapidly, data accumulates in a speed unmatchable by human capacity of data processing[2]. Data mining as a multidisciplinary joint effort from databases, machine learning and statistics, is championing in turning mountains of data into nuggets. Researchers and practitioners realize that in order to use data mining tools effectively, data processing is essential to successful data mining.PrimitiveThese are features which have an influence on the output and their role cannot be assumed by the rest.[1].Feature selection can be found in many areas of data mining such as classification, clustering, association rules and regression. For example, feature selection is called subset or variable selection in statistics. In this survey, we focus on feature selection algorithms for classification. Early research efforts mainly focus on feature selection for classification with labelled data. (Supervised feature selection) where class information available. The latest developments however show that the above general procedure can be well adopted to feature selection for classification with unlabeled data,(or unsupervised feature selection) where data is unlabeled.

Feature selection is the process of selecting the best feature among all the features because all the features are not useful in constructing the clusters: some features may be redundant or irrelevant thus not contributing to the learning process.Feature selection, a process of choosing a subset of features from the original ones, is frequently used as a preprocessing technique in data mining. It has proven effective in reducing dimensionality, improving mining efficiency, increasing mining accuracy, and enhancing result comprehensibility. Feature selection methods can broadly fall into the wrapper model and the filter model.The wrapper model uses the predictive accuracy of a

predetermined mining algorithm to determine the goodness of a selected subset. It is computationally expensive for data with a large number of features. The filter model separates feature selection from classifier learning and relies on general characteristics of the training data to select feature subsets that are independent of any mining algorithms [1]. By reducing the number of features, one can both reduce over fitting of learning methods, and increase the computation speed of prediction. In this paper on the selection of a few features among several in a context of classification. Main interest in this paper is to design an efficient filter, both from a statistical and from a computational point of view.

Feature selection algorithms designed with different evaluation criteria broadly fall into three categories: the Filter model, the Wrapper model, and the Hybrid model [2]. The Filter model relies on general characteristics of the data to evaluate and select feature subsets without involving any mining algorithm. The Wrapper model requires one predefined mining algorithm and uses its performance as the evaluation criterion.

II. Proposed Algorithm

J48 is slightly modified C4.5 decision tree for classification. The C4.5 algorithm generates a classification decision tree for the give data set by recursive partitioning of data. The decision is grown using depth first strategy (DFS). The algorithm considers all the possible tests that can split the data set and selects a test that gives the best information gain. The formulations of J48 are proposed by and described below.

Input: T is training data

Output: A decision tree.

1. If (**T** belong to the same category **C**) then **Return N** as a leaf node, and mark it as class **C**;
2. If attribute is the remainder samples of **T** is less than a give value, then Return **N** as a leaf node, and mark it as the category which appears most frequently in attribute, for each attribute, calculate its information gain ratio.
3. Suppose attribute is the testing attribute of **N**, the test attribute equal to the attribute which has the highest information gain ratio in attribute list.
4. If testing attribute is continuous, the find its division threshold.
5. For each new leaf node grown by node **N**
{
Suppose **T** is the sample subset corresponding to the leaf node. If **T** has only a decision category, then mark the leaf node as this category; else continue to implement J48-Tree
}
6. Compute the classification error rate of each node, and then prune the tree.

III. System Implementation of Proposed System

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. The Pre-process panel has facilities for importing data from a database, and for pre-processing this data using a filtering algorithm. These filters can be used to transform the data and make it possible to delete instances and attributes according to specific criteria.

Confusion matrix

A confusion matrix contains information about actual and predicted classifications done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix. The following table shows the confusion matrix for a two class classifier.

The entries in the confusion matrix have the following meaning in the context of our study:

1. a is the number of correct predictions that an instance is negative,
2. b is the number of incorrect predictions that an instance is positive,
3. c is the number of incorrect of predictions that an instance negative, and
4. d is the number of correct predictions that an instance is positive.

IV. Experimental Results and 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, 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 datasets 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

Table 2: Classification Accuracy of Different Datasets.

S.NO	Dataset	Decision Tree Accuracy	J48
1	cmc	54.99%	52.14%
2	Hepatitis	81.29%	83.87%
3	Ionosphere	89.46%	91.45%
4	labor	77.19%	73.68%
5	lung-cancer	75%	78.13%
6	Mushroom	100%	100%
7	pima_diabetes	73.31%	73.83%
8	Sponge	92.11%	92.11%
9	Spambase	91.41%	92.98%
10	Vehicle	65.01%	72.46%
11	Waveform	73.80%	75.08%
12	Zoo	91.09%	92.08%
13	Nursery	94.69%	97.05%
Average Accuracy		81.48%	82.68%

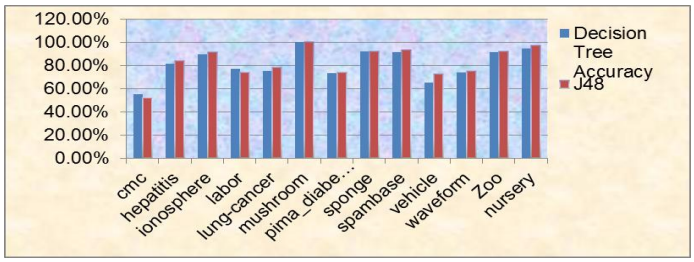

Figure 1: Performance Analysis graph comparing of Decision Tree & J48

Table 3: Classification Accuracy of J48, IBK, Naive Bayes, Decision Tree

S.No	Name of the Dataset	J48	IBK	Naïve Bayes	Decision Tree Accuracy
1	cmc	52.14%	44.33%	50.78%	54.99%
2	hepatitis	83.87%	80.65%	84.52%	81.29%
3	ionosphere	91.45%	86.32%	82.62%	89.46%
4	labor	73.68%	82.46%	89.47%	77.19%
5	lung-cancer	78.13%	68.75%	78.13%	75%
6	mushroom	100%	100%	95.83%	100%
7	pima_diabetes	73.83%	70.18%	76.30%	73.31%
8	sponge	92.11%	92.11%	92.11%	92.11%
9	spambase	92.98%	90.78%	79.29%	91.41%
10	vehicle	72.46%	69.86%	44.80%	65.01%
11	waveform	75.08%	73.62%	80%	73.80%
12	Zoo	92.08%	96.04%	95.05%	91.09%
13	nursery	97.05%	98.38%	90.32%	94.69%
Average of Accuracy		82.68%	81.03%	79.93%	81.48%

Table 4: No. of Features selected for each dataset with different feature selectors

S.NO	Name of the Dataset	CfsSubSetEval	ChiSquaredAttributeEval	ConsistencySubsetEval	GainRatioAttributeEval	InfoGainAttributeEval	OneRAttributeEval	Principal Components	ReliefFAttributeEval	SymmetricalUncertAttributeEval
1	cmc	3	9	9	9	9	9	15	9	9
2	hepatitis	10	19	12	19	19	19	16	19	19
3	ionosphere	14	34	7	34	34	34	23	34	34
4	labor	4	16	4	16	16	16	16	16	16
5	lung-cancer	8	56	4	56	56	56	25	56	56
6	mushroom	1	22	5	22	22	22	59	22	22
7	pima_diabetes	3	8	8	8	8	8	8	8	8
8	sponge	3	45	1	45	45	45	65	45	45
9	spambase	10	57	25	57	57	57	48	57	57
10	vehicle	11	18	18	18	18	18	7	18	18
11	waveform	15	40	11	40	40	40	34	40	40
12	Zoo	10	17	1	17	17	17	94	17	17
13	nursery	1	8	8	8	8	8	18	8	8

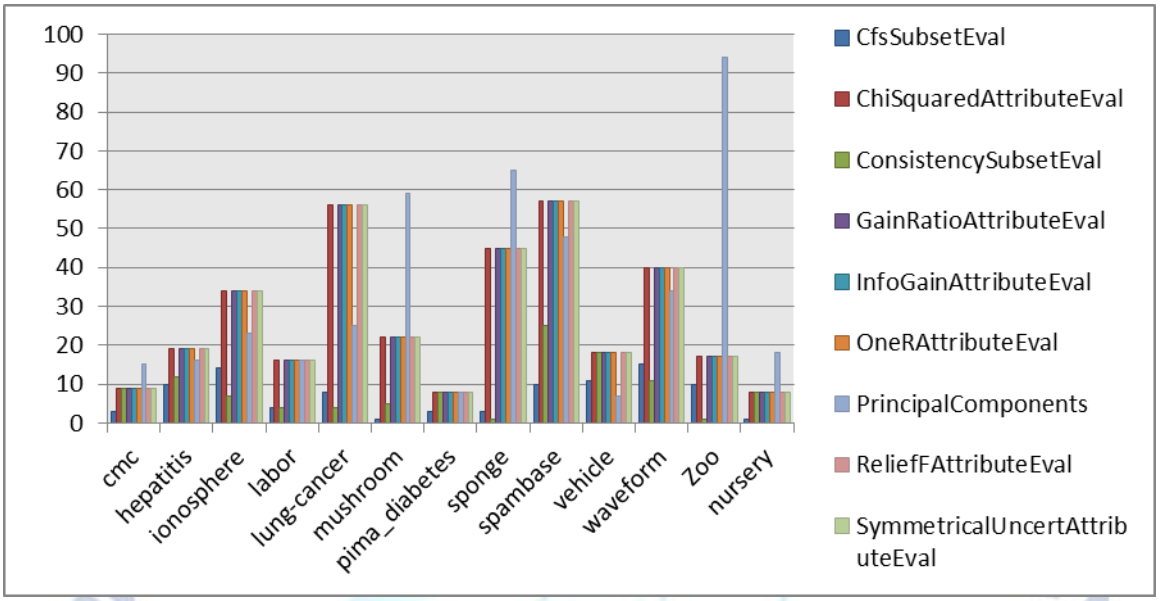

Figure 2: Performance Graph for All feature selectors with datasets

V. CONCLUSION

This paper proposes a J48 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. In this average accuracy of J48 Classification algorithm gives better performance comparing to other algorithms.

REFERENCES

[1] H. Liu, H. Motoda, R. Setiono, Z. Zhao, "Feature Selection: An Ever Evolving Frontier in Data Mining", JMLR: Workshop and Conference Proceedings 2010, Volume: 4, Publisher: Cite seer, pages 4-13.

[2] L. Yu, H. Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation- Based Filter Solution", Proceedings of the Twentieth International Conference on Machine Leaning, ICML-03, Washington, D.C., August, 2003, pages 856-863.

[3] A.Veeraswamy, Dr.S.Appavu alias Balamurugan, "A Survey of Feature Selection Algorithms in Data mining" Proceedings of the "3rd International Conference on Trendz in Information Sciences and Computing" (TISC-2011) Organized by Sathyabama University, Chennai, India, 8th & 9th December 2011, Pages 40-46.

[4] 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.

[5] Sombut Foithong, Ouen Pinngern, Boonwat Attachoo," Feature subset selection wrapper based on mutual information and rough sets" , Expert Systems with Applications 39 (2012) 574–584.

[6] Qinbao Song, Jingjie Ni and Guangtao Wang," A Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data" IEEE Transactions on Knowledge and Data Engineering Vol: 25 No: 1 Year 2013.

[7] A.Veeraswamy, Dr.S.Appavu alias Balamurugan, Dr.E.Kannan, "Evaluation of Feature selection method for Classification of Data Using Support Vector Machine Algorithm Springer International Publishing Advances in Intelligent Systems and Computing in ISBN: 978-3-319-03107.

[8] A.Veeraswamy, Dr.S.Appavu alias Balamurugan, Dr.E.Kannan "An Implementation of Efficient Data mining Classification Algorithm using NBTREE" International Journal of Computer Applications (0975 – 8887) Volume 67– No.12, April 2013.

[9] Yonghong Peng, Zhiqing Wu, Jianmin Jiang," A novel feature selection approach for biomedical data classification", Journal of Biomedical Informatics 43 (2010) 15–23.

[10] Song, Q., Ni, J. and Wang, G. (2013) A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data. IEEE Transactions on Knowledge and Data Engineering, 25, 1-14.

[11] Wald R. , Khoshgoftaar T. M. , Napolitano A. , "Stability Of Filter- And Wrapper-Based Feature Subset Selection", IEEE 25th International Conference On Tools With Artificial Intelligence, pp. 374 – 380,2013.