ISSN 2250-3153

Implementation of decision tree algorithm c4.5

[1]Harvinder Chauhan, [2]Anu Chauhan

[1]Assistant Professor, P.G.Dept. of Computer Science
Kamla Nehru College ForWomen, Phagwara (Punjab)
harrymit21@yahoo.com

[2]Research Scholar
Anu.chauhan711@yahoo.com

Abstract-Data classification is a form of data analysis that can be used to extract models describing important data classes. There are many classification algorithms but decision tree is the most commonly used algorithm because of its ease of implementation and easier to understand compared to other classification algorithms**.**C4.5 is one of the most effective classification method. In this paper we are implementing this algorithm using weka data mining tool using publicly available datasets of different size. This paper also gives insights into the rate of accuracy it provides when a dataset contains noisy data, missing data and large amount of data.

I. INTRODUCTION

In the past, to extract information by data analysis was a manual and pain staking process because much domain knowledge was required, and understanding of statistical approach is also needed. However such approach will become inappropriate while facing the rapidly growing sizes and dimensions of the data. A community of researchers devoted themselves to the field called "data mining" to solve automating data analysis problem and discover the implicit information within the huge data (Giordana and neri,1995).Data classification is one of data mining techniques used to extract models describing important data classes. Some of the common classification methods used in data mining are: decision tree classifiers, Bayesian classifiers, k-nearest-neighbor classifiers, case-based reasoning, genetic algorithms, rough sets, and fuzzy logic techniques. Among these classification algorithms decision tree algorithms is the most commonly used because of it is easy to understand and cheap to implement. Most Decision tree algorithms can be implemented in both serial and parallel form while others can only be implemented in either serial or parallel form. Parallel implementation of decision tree algorithms is desirable in-order to ensure fast generation of results especially with the classification/prediction of large data sets, it also exploits the underlying computer architecture (Shafer et al, 1996). But serial implementation of decision algorithm is easy to implement and desirable when small-medium data sets are involved. In this paper we will implement c4.5 the most common decision tree algorithm using weka, serially.

II. IMPLEMENTATION OF C4.5 ALGORITHM

In order to classify our data, first we need to load the dataset. This will be done in wekaexplorer window. Here, we have loaded wheather dataset having 14 instances and 4 attributes. On the basis of information contained in this dataset, weka enable us to make a decision whether or not to play a particular game on the basis of the wheather conditions. As shown in the figure given below weka explorer contains various tabs at the top of the window. User can choose one of them according to his task.

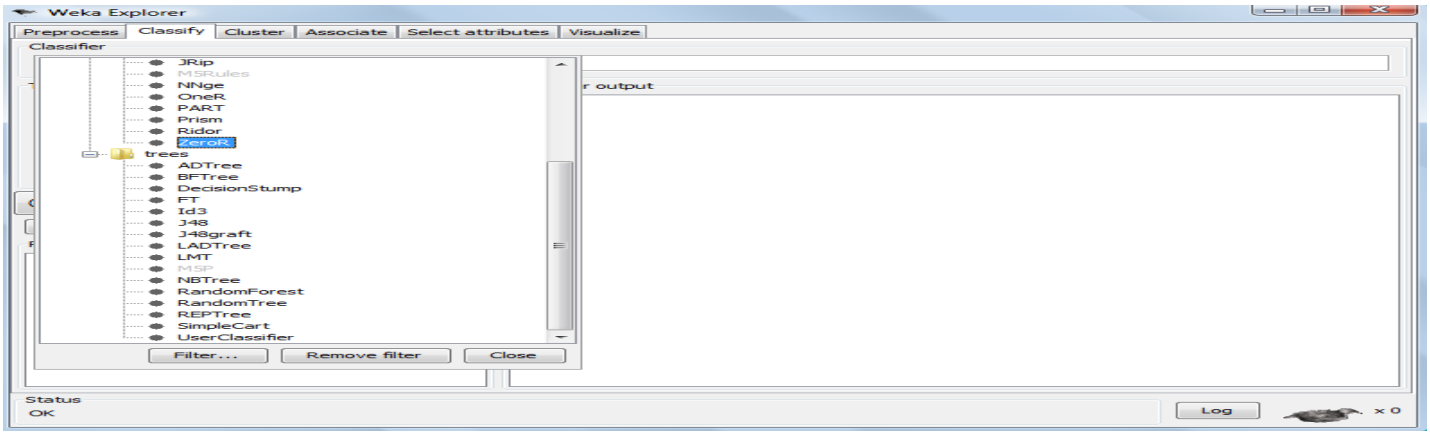

Figure1: classification panel

As per our research is concerned we need to click on classify tab. This window consists of various classifiers like bays, functions, lazy, meta and tree etc. available in weka. We first click on trees, then choose J48 (c4.5 is termed as J48 in weka software) which results in following figure.

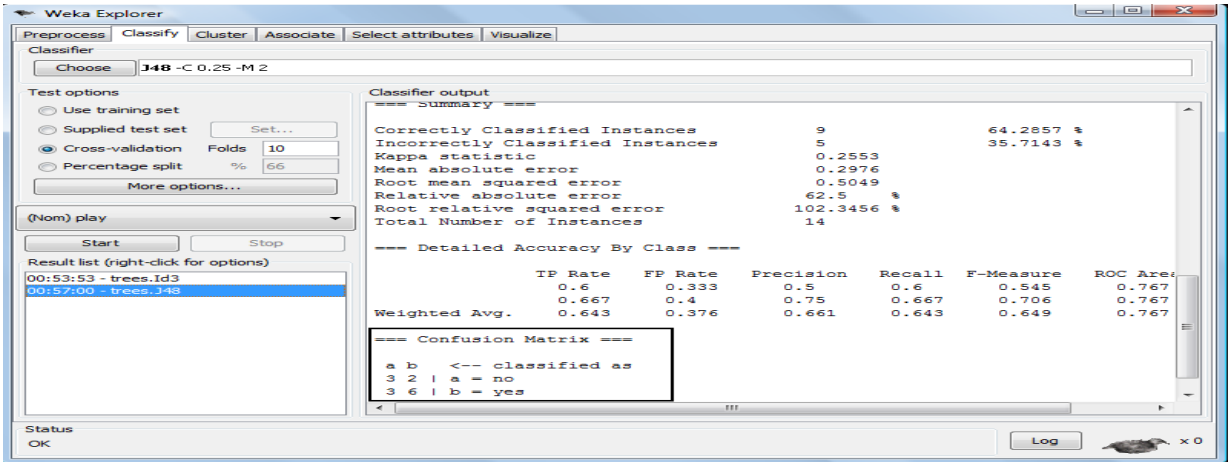

Figure2: Weka run information for C4.5

The name of the classifier listed in the text box right beside the choose button. The text after j48 represents parameter setting. These are default parameters which states that the confidence factor for pruning is 0.25, to use binary splits and restrict the minimum no. of instances in a leaf to 2 which means grow the tree fully. We can set these parameters according to our convenience too. After setting parameters we choose one of the test methods as four are given here. Finally click on the start button to build decision tree. The right hand side window shows classifier output, which shows some text and numeric values which is not easy to interpret. We will discuss it later. So let us look at graphical representation of this tree by choosing visualize tree option. In this way weka grow decision tree applyingc4.5 algorithm of data classification technique in data mining.

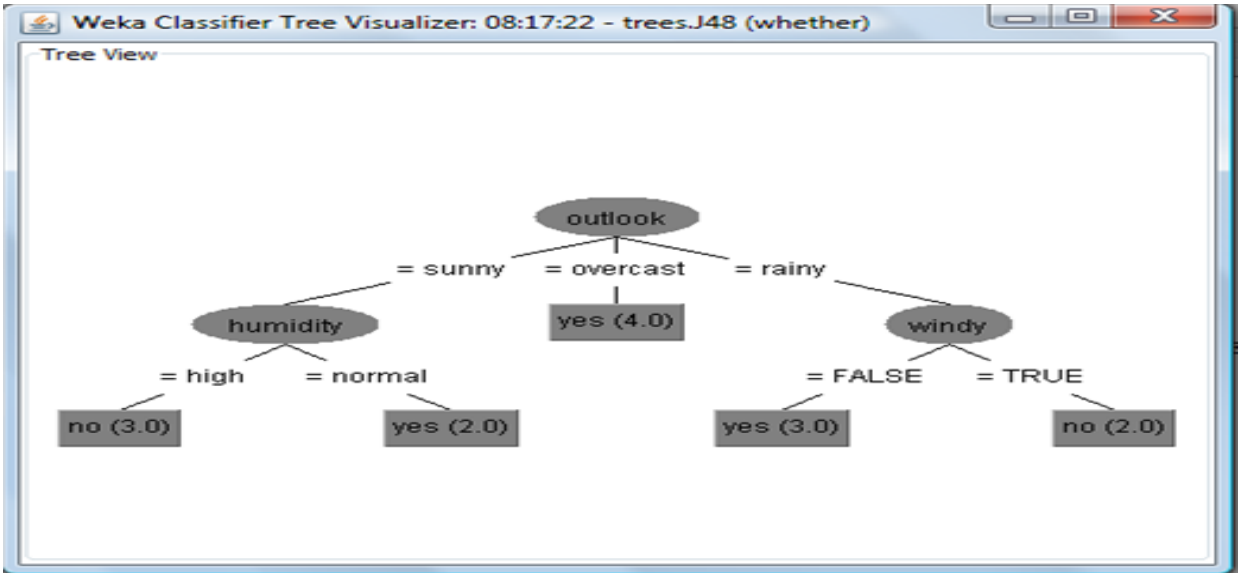

Figure3:Graphical representation of decision tree using C4.5

III. ACCURACY EVALUATION

The accuracy of a classifier is the percentage of test set tuples that are correctly classified by the classifier. There are four accuracy evaluation methods in weka, termed as test methods. In this paper we have used 10 fold cross validation method due to its low biasness. For estimating the accuracy of c4.5 we have taken different datasets of different size.

3.1 Effect of noisy data

Data that contains errors due to human mistakes, expert's misjudgment in classifying training examples etc. When noisy datasets are applied to c4.5 algorithm, it still provides greater accuracy as it employs tree pruning methods which avoid noise from data. The obtained accuracy rates are shown in the table below.

Table1: Effect of noise

Dataset	Accuracy rate (%)
Bank_class	65
Iris	86
Cardata	82
Soybean	83

3.2 Effect of missing data

Applying a dataset having some missing values in it results in producing good results. No doubt noise effects its performance but there are certain decision tree algorithm that are not even able to handle missing values in a dataset.

Following table shows the results obtained by applying missing values to c4.5. This table signifies that as the rate of missing values increases the rate of accuracy starts falling down.

Table2: Effect of missing values

Missing value rate (%) for dataset iris	Accuracy rate (%)
4	96%
16	94%
20	90%

3.3Effect of scalability

Finally c4.5 is applied with large datasets to check how efficiently it performs with large amount of data. It provides good accuracy in this case also. For this we have taken 2 datasets of varying size.

Table3: Effect of scalability

Datasets	Instances	Accuracy rate
Cardata	1728	92%
mushroom	2074	88%

IV. Conclusions and lights to the future

In this paper we first show implementation of c4.5 decision tree algorithm. After that rate of accuracy it provides when dataset contains noise, when there is some missing data in a dataset and when a dataset contains number of instances in it. The experimentalresults show that c4.5 provides greater accuracyin each above said case.In this study we focused on serial implementation of decision tree algorithm which ismemory resident, fast and easy to implement. In future we will go for its parallel implementation which is comparatively complex and evaluate how much accuracy this algorithm provides in that case.

References

1. Anyanwu, M., and Shiva, S. (2009). Application of Enhanced Decision Tree Algorithm toChurn Analysis. 2009 International Conference on Artificial Intelligence and PatternRecognition (AIPR-09), Orlando Florida

2. Jiawei han and micheline kamber. Data mining concepts and techniques, second edition,285-291

3. Matthew N.anyanwu, Sajjan g. shiva. Comparative analysis of serial decision tree classification algorithms

4. Mehdi piroozma, Youping deng, jack y yang and mary qu yang. A comparative study of different machine learning methods on microarray gene expression data, BMC genomics

5. Tzung-I tang,Gang Zheng, Yalou huang,Guangfu Shu,Pengtao wang. A comparative study of medical data classification methods based on decision tree and system reconstruction analysis. IEMS vol.4,no.1,pp-102-108,june 2005

6. Xu,M, wang, J. Chen, T. (2006). Improved decision tree algorithm: ID3+, intelligent computing in signal Processing and pattern recognition, Vol. 345, PP.141-149