Computer Engineering and Applications Vol. 4, No. 3, September 2015

C4.5 Versus Other Decision Trees: A Review

Salih ÖZSOY [1], Gökhan GÜMÜŞ [2], Savriddin KHALILOV[3]
IT Department, Ishik University, 100 Meter St., Erbil, Iraq
salih.ozsoy@ishik.edu.iq[1], gokhan.gumus@ishik.edu.iq[2], savriddin.halil@ishik.edu.iq[3]

ABSTRACT

In this study, Data Mining, one of the latest technologies of the Information Systems, was introduced and Classification a Data Mining method and the Classification algorithms were discussed. A classification was applied by using C4.5 decision tree algorithm on a dataset about Labor Relations from http://archive.ics.uci.edu/ml/datasets.html. Finally, C4.5 algorithm was compared to some other decision tree algorithms. C4.5 was the one of the successful classifier.

Keywords: Data mining, Classification, C4.5 algorithm, Decision tree

1. INTRODUCTION

Information has been always valuable for the mankind. The ages before the Sumerian invented writing are considered as Dark Ages. Along with the invention of writing information could be recorded and transferred from generation to generation.

Nowadays computers are used to create, store and transfer the information. Not only computers but also data communication technologies are developed. Moreover, one of the notable important points is the products based on technology becomes increasingly cheaper. End users could own faster and more skilled computers. Nowadays data is currently stored in digital mediums. The capacities of storage mediums are increased unlike the prices.

Today's Information problem is quite different than the past. Now, it is the fact that data is accessible as much as you do not need. Data Mining could be defined as a technology to distinguish the valuable data from worthless data and present as useful information. In brief data mining can be defined as; to figure out a pattern from dataset by using an application, which has on purpose algorithms .

2. METHODOLOGY

There were many methods and algorithms developed for Data mining. Most of the Data Mining methods are statistical based. There are many Data mining methods and techniques and the method is chosen according to the definition of the problem and the structure of data. That's why it is not possible to mention a best method or algorithm.

This study focused on classification. So only classification method was discussed as well as some classification algorithms were compared.

2.1 DATA CLASSIFICATION

It is possible to classify data by using common features. For instance, a company can classify their customers by considering some features or habits. A super market administration may want to classify the customers according to their order dues. The

customers that have order dues less than the average could be classified as 'Ordinary' and the customers that have order dues more than average could be classified as 'Wealthy'.

Similarly it is possible to make classification by revealing common features or differences in a dataset. Classification is based on a learning algorithm. Whole dataset is not but a part is used for training. Goal of the learning is having a classification model. In other words, Classification is the process of determining the classes of the instances whose classes are unknown.

For example, the customers could be roughly classified into 2 groups: 'Those who pay on time' and 'those who don't pay on time'.

Classification is used in a variety of fields. For instance, classifications of the trends in financial markets or evaluating a credit demand in a bank.

Process of classification of data consists of 2 steps.

Step 1: Building the Model
Step 2: Applying the Model

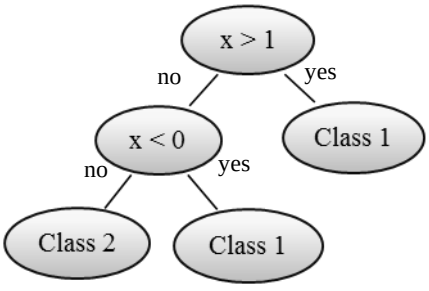
First step is building a model by using existing data. This model is built according to the attributes which each of are a particular feature of the observations in the dataset. Some of the instances in the dataset are used to build this classifier model. Second step is applying this model. After determining the rules for the classification, these rules will be tested on the new data to get desired results.

2.2 DECISION TREES AND CLASSIFICATION

One of the approaches to classify the dataset is named as decision trees. Accordingly applicable statistics different decision tree algorithms were developed under the name of machine learning. There are many ways of learning methods, which uses predefined sets for building up a decision tree.

Decision trees repeat themselves for distributing the number of data into sub-groups, this iterative progress repeats until all the data has been grouped accordingly desired condition. When the items in the dataset divided into groups, each group member represents more common features. So the relations among homogenous sets can be realized and evaluated by the produced figure[i].

Decision trees looks like flowchart diagrams. Each attribute is demonstrated by a nod. Branches and the leaves are members of tree structure. The ultimate structure is called as "leave", the top most structure is called as "root" and the structures in between these two are called as "branch". Figure 1 demonstrates a common look of a decision tree. Decision trees offer suitable infrastructure for applying classification.



ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

FIGURE 1. A basic decision tree

The figure 1 demonstrates a basic decision tree, which was built upon "X" from a dataset. Related to figure 1, X is a real number. Class 1 indicates those numbers, which are smaller than their square and class 2 represents those numbers, which are greater than their square. Accordingly figure 1 if x is greater than 1 that means number belongs to class 1. If x is not greater than 1 and it is also less than 0 again the number classified as class 1. Finally if x is not greater than 1 and also not less than 0 the number became a member of class 2.

Data miners have been continuously developing many decision tree algorithms to classify data. For instance, researchers from Poland propose a new algorithm, which is based on the commonly known CART algorithm called the dsCART algorithm in 2014.[ii]

Researchers from China have compared the impacts of the missing data toleration technique of C4.5 with the k-NN missing data imputation method on the prediction accuracy of C4.5 in the context of software cost prediction in 2008. They found that k-NN imputation can improve the prediction accuracy of C4.5 and the improvements are statistical significant and both C4.5 and k-NN can be affected by the missingness mechanism, the missing data pattern and the missing data percentage.[iii]

Some researchers studied the classification of blood characteristics by a C4.5 decision tree, a naïve Bayes classifier and a multilayer perceptron for thalassaemia screening. Their aim was to classify eighteen classes of thalassaemia abnormality, which have a high prevalence in Thailand, and one control class by inspect- ing data characterised by a complete blood count (CBC) and haemoglobin typing in 2011. Their experiment involving stratified 10-fold cross-validation revealed that both naïve Bayes classifier and multilayer perceptron are the most suitable classifier for the data that has been pre-processed by attribute discretisation.[iv]

Researchers from India used C4.5 Algorithm on a web based Soya Bean Expert System in 2012. The Proposed Bagging algorithm was used to improve the performance of C4.5. Researchers' approach could improve the performance of C4.5 between 4% to 6%.[v]

Carlos J. Mantas, Joaquín Abellán have presented a new model called Credal-C4.5, a modified version of the C4.5 algorithm. It has been defined by using a mathematical theory of imprecise probabilities and uncertainty measures on credal sets. They have showed, C4.5 and Credal-C4.5 are very similar in performance when no noise is added, and the only difference is that Credal-C4.5 presents trees with a notable lower number of nodes. When noise is added, Credal- C4.5 has a better performance than C4.5, and, in this case, also the number of nodes of Credal-C4.5 is notably lower.[vi]

The one of the major problems of decision trees is "What should be the criterion that leads the splitting of roods or managing to be branched?". Actually there is a possible decision tree algorithm can be generated for each criteria. In this study, C4.5 was applied. C4.5 is using entropy-based segmentation algorithm and it is used widely for designing decision trees[vii]. Eventually C4.5 algorithm was also compared against ADTree, BFTree, DecisonStump, FT, LADTree, LMT, NBTree, RandomForest, RandomTree, REPTree and SimpleCart.

2.3 APPLICATION

In this study a dataset about Labor Relations from http://archive.ics.uci.edu/ml/datasets.html web page was classified by using Weka Application.

Dataset contains real-life data, which was shared by Ottova University by 1988. The information about dataset shown in table 1:

TABEL 1.

Dataset Information

Characteristics of Dataset	Multivariate	Number of Instances:	57
Characteristics of Attributes	Categorized, integer, real number	Number of Attributes:	16

Detailed information can be found on
https://archive.ics.uci.edu/ml/datasets/Labor+Relations

For this study, Weka, a Java based application which was developed by New Zealand Waikato University was used. Weka 3.6.8 was selected due to; being a basic application at data mining, having no restriction for data size and a rich content about modeling, being a free product, analyzing the quality of data and to evaluate the data visually[viii]. The original dataset contains 16 attributes but it looks like there are 17 attributes. The reason is the class attribute, which evaluates the agreement either good or bad.

Before this progress, a graph in red and blue colors is located at the right bottom of user interface. In this graph columns represent different values of selected attribute, blue color indicates class of bad and red color indicates the class of good agreements. For example duration of the agreement has 3 different values, therefore there are three columns drawn. For instance 5 good and 5 bad agreements so totally10 conditions are present about those agreements which has 1 as duration, likewise 10 bad and 17 good agreement, in total 27 conditions were grouped under the duration as 2, finally 19 conditions which are 5 bad and 12 good agreements collected which has duration as 3. When user clicks on “Visualize all” button, application draws related graphs for all attributes. The graph of all attributes was illustrated at figure 2.

2.4 LOADING THE DATASET TO THE APPLICATION

For this study, Weka, a Java based application which was developed by New Zealand Waikato University was used. Weka 3.6.8 was selected due to; being a basic application at data mining, having no restriction for data size and a rich content about modeling, being a free product, analyzing the quality of data and to evaluate the data visually. The original dataset contains 16 attributes but it looks like there are 17 attributes. The reason is the class attribute, which evaluates the agreement either good or bad.

Before this progress, a graph in red and blue colors is located at the right bottom of user interface. In this graph columns represent different values of selected

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

attribute, blue color indicates class of bad and red color indicates the class of good agreements. For example duration of the agreement has 3 different values, therefore there are three columns drawn. For instance 5 good and 5 bad agreements so totally10 conditions are present about those agreements which has 1 as duration, likewise 10 bad and 17 good agreement, in total 27 conditions were grouped under the duration as 2, finally 19 conditions which are 5 bad and 12 good agreements collected which has duration as 3. When user clicks on "Visualize all" button, application draws related graphs for all attributes. The graph of all attributes was illustrated at figure 2

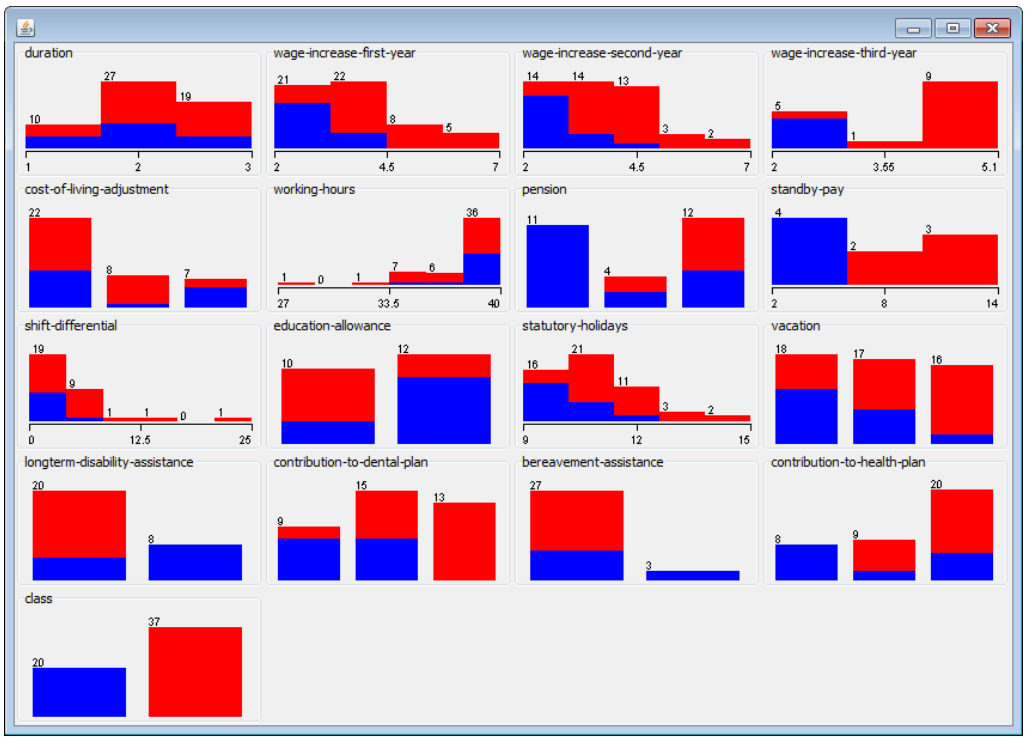


FIGURE 2. Automatic demonstration of all attributes in terms of good and bad by Weka

2.5 SELECTING ALGORITHM FOR APPLICATION

As it has shown at figure 2 the dataset contains numeric values and some attributes contains missing data. C4.5 can generate decision trees by numeric values. Beside it also offers a solution to build decision trees when there are missing values. Due to discussed points, C4.5 algorithm was selected to build decision tree.

2.6 MODELING THE CLASSIFIER

From the test options 50 was selected for divide by percentage, by this way 57 instances were divided into two groups, 29 of them used for teaching set and 28 of them used for test set. After instances were grouped classifier get trained by teaching set and modeled.

After the previous progress "Visualize Tree" option from the right click menu was selected. Eventually the decision tree of decision rules had been generated

as given in figure 3 by Weka. C4.5 algorithm automatically ignores the irrelevant variables and sets the variables of new learning progress. The major reason why some variables were ignored is; the correlation in between the variables is low and some variables remain lower than the correlation coefficient. Correlation coefficient represents the direction and the degree of relation in between variables. After the training, Weka generates a pruned-tree structure which has 3 leaves.

Rules are as mentioned below;
Rule1: if wage increase first year <=2.5
Agreement is bad
Rule2: if wage increase first year >2.5
And statutory-holiday <=10
Agreement is bad
Rule3: If wage increase first year >2.5
And statutory-holiday >10
Agreement is good

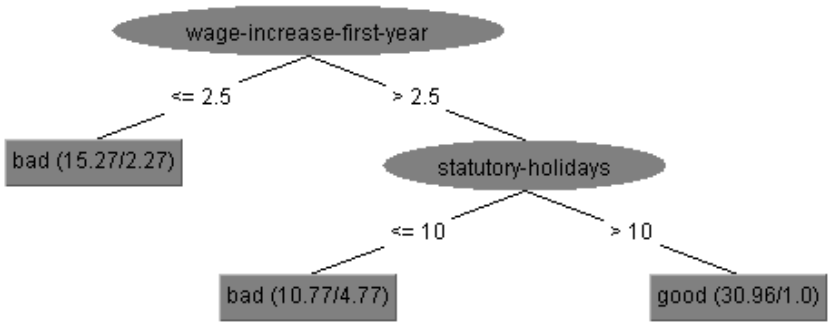


FIGURE 3. Decision tree

2.7 TESTING THE CLASSIFIER

The classifier was trained and built on the 29 instances. Then classifier was tested with the rest 28 instances. At the end of the test, 24 of the instances were classified correctly which means that the performance of the classifier is %85.71.

If we look at the the confusion matrix, 7 of the bad class among 9 were classified correctly and 17 of the good class among 19 were classified correctly. 2 wrong classification for each class has been done.

As mentioned before, the classification performance is %85.71. in case of dividing the dataset as %50 training and %50 test sets. In this section the performance comparision was discussed up on the changes in percentages of training and the test sets.

First of all we divided our dataset into %25 training and %75 test set. İt was observed that the same decision rules and the decision tree were generated. That means the classifier was modeled as exactly same with 14 instances instead of 29. When we test the classifier with the 43 instances, we observed that 32 instances classified correctly. That means the classification performance in this case was %78.05.

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

The reason of the reduce in performance is not training the classiffier enough. Because the same decision rules and the tree were generated. It is thought that the quantity increasement of the instances in test cause the reduce in performance.

After this result, we splitted up the %75 of the instances as training and %25 of the instances as test sets. After the training of the classifier with 43 instances, the same decision rules and the tree were generated as we expected. Then we tested the classifier with the rest 14 istances. 11 of the instances were classified correctly, and 3 of them classified incorrectly. That was %78.57 correct classification. This test's performance was also worse than the first one. So it is understood that increasing the quantity of instances for training does not work to increase the performance of the classification. It is thought that the reduce in performance is due to coinciding of 3 challenging instances.

TABLE 1.

Performance Changes with different Divison of the Dataset

Division	True Classification	False Classification	Performance
%25 Training - %75 Test	32	9	78.05%
%50 Training - %50 Test	24	4	85.71%
%75 Training - %25 Test	11	3	78.57%

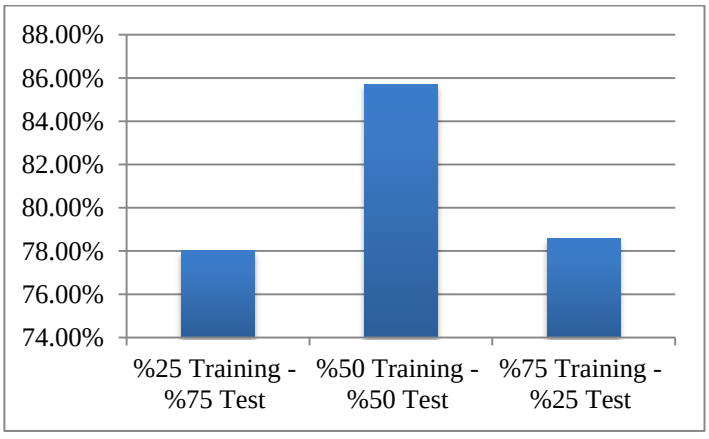


FIGURE 4. %50 Training and %50 Test Sets work fine.

Test is tried with several training and test ratios. But the same desicion rules and tree was got. So it was concluded that optimizing the classifier with this way is not possible.

2.8 PERFORMANCE COMPARISION OF THE CLASSIFIER

In this section C 4.5 algorithm was compared to other tree classifiers which are shown in table 2 and figure 5. In this section training and test sets were divided by %50 for each classification. So each classifier was trained with the first 29 of the instances and tested with the rest 28 instances.

Dataset contains some quantitative attributes and some of the values are missing. That's why ID3 algoritm, which C 4.5 is based on could not classify the dataset. The best classification performance with 25 true classification out of 28 instances belonged to Random Forest among 13 classifiers. J48 classifier that uses C4.5 algorithm could classified 24 of the instances correctly and shared the second place with the other 7 classifers. 3 of the classifiers had worse performace than C 4.5. Comparision of the performances of the classifiers could be seen on table 2 and figure 5.

TABLE 2.
Performance Comparision of Tree Classifiers

Classifier	True Classification	False Classification	Performance
ADTree	24	4	85.71%
BFTree	24	4	85.71%
DecisionStump	23	5	82.14%
FT	19	9	67.86%
J48 (C4.5)	24	4	85.71%
LADTree	24	4	85.71%
LMT	24	4	85.71%
NBTree	24	4	85.71%
Random Forest	25	3	89.29%
Random Tree	24	4	85.71%
REPTree	23	5	82.14%
SimpleCart	24	4	85.71%

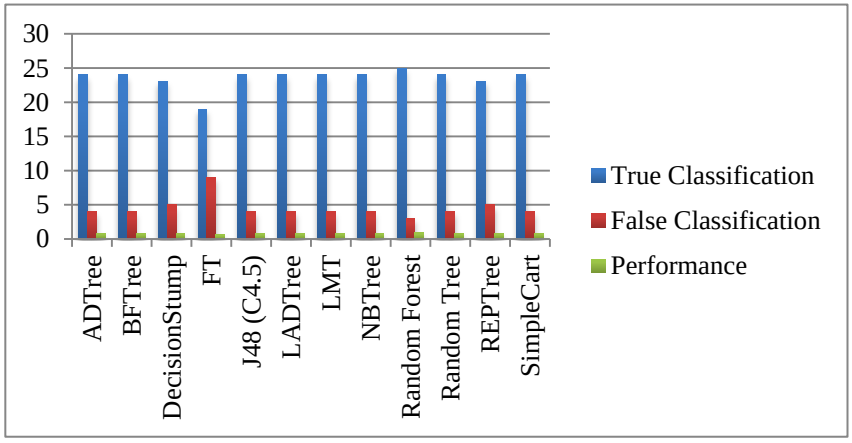


FIGURE 5. Performance Chart

3. CONCLUSION

C4.5 algorithm makes possible to classify the datasets that has quantitative attributes. In addition it is possible to handle missing values with this algorithm. Handling both continuous and discrete attributes. In order to handle continuous attributes, C4.5 creates a threshold and then splits the list into those whose attribute value is above the threshold and those that are less than or equal to it.[ix] Used dataset in this study is absolutely real and is about local labor agreements in Canada. According to the values of the attributes, each agreement is classified as acceptable

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

or unacceptable. First of all C 4.5 algorithm was used on dataset with different Training and Test sets Division ratios. It was observed that changing the ratio does not effect the decision rules and tree. So it is concluded that this way can not improve the performance. When we compare C 4.5 algorithm with the other tree classifiers, it was only worse than Random Forest. C 4.5 had a pretty good performance on such a dataset that has many quantitative varibles and missing values.

REFERENCES

[1] Yadav, Malik, Chandel , Selection of most relevant input parameters using WEKA for artificial neural network based solar radiation prediction models, Renewable and Sustainable Energy Reviews, 2014.

[2] Leszek Rutkowski, Maciej Jaworski, Lena Pietruczuk, Piotr Duda, he CART decision tree for mining data streams, Information Sciences 266, p1–15, 2014.

[3] Qinbao Song a, Martin Shepperd, Xiangru Chen, Jun Liu, Can k-NN imputation improve the performance of C4.5 with small software project data sets? A comparative evaluation, The Journal of Systems and Software 81, 2361–2370, 2008.

[4] Damrongrit Setsirichok, Theera Piroonratana, Waranyu Wongseree, Touchpong Usavanarong, Nuttawut Paulkhaolarn, Chompunut Kanjanakorn, Monchan Sirikong, Chanin Limwongse, Nachol Chaiyaratan, Classification of complete blood count and haemoglobin typing data by a C4.5 decision tree, a naïve Bayes classifier and a multilayer perceptron for thalassaemia screening, Biomedical Signal Processing and Control, 7, 202–212, 2012

[5] M.S. Prasad Babu, Swetha Reddy, B. Venkata Ramana, N.V. Ramana Murty, A Web-Based Soya Bean Expert System Using Bagging Algorithm with C4.5 Decision trees, International Journal of Agriculture Innovations and Research, Volume 1, Issue 4, ISSN (Online) 2319-1473, 2012.

[6] Carlos J. Mantas, Joaquín Abellán , Expert Systems with Applications, 4625–4637, 2014.

[7] Joos, P., Vadhoof, K., Ooghe, H., Sherens, N., Credit classification: a comparison of logit models and decision trees. Proceedings Notes of the Workshop on Application of Machine Learning and Data Mining in Finance. 10th European Conference on Machine Learning, p59–72, 1998.

[8] Fayyad, U., Shapiro, G.P., Smyth, P., From data mining to knowledge discovery in databases. AI Mag. 17 (3), p37–54, 1996.

[9] J. R. Quinlan. Improved use of continuous attributes in c4.5. Journal of Artificial Intelligence Research, 4, p77-90, 1996.