Improved C4.5 Algorithm for Rule Based Classification

MOHAMMED M MAZID, A B M SHAWKAT ALI, KEVIN S TICKLE
School of Computing Science
Central Queensland University
AUSTRALIA.
E-mail: {m.mazid, s.ali, k.tickle@cqu.edu.au}

Abstract: - C4.5 is one of the most popular algorithms for rule base classification. There are many empirical features in this algorithm such as continuous number categorization, missing value handling, etc. However in many cases it takes more processing time and provides less accuracy rate for correctly classified instances. On the other hand, a large dataset might contain hundreds of attributes. We need to choose most related attributes among them to perform higher accuracy using C4.5. It is also a difficult task to choose a proper algorithm to perform efficient and perfect classification. With our proposed method, we select the most relevant attributes from a dataset by reducing input space and simultaneously improve the performance of this algorithm. The improved performance is measured based on better accuracy and less computational complexity. We measure Entropy of Information Theory to identify the central attribute for a dataset. Then apply correlation coefficient measure namely, Pearson's, Spearman, Kendall correlation utilizing the central attribute of the same dataset. We conduct a comparative study using these three most popular correlation coefficient measures to choose the best method on eight well known data mining problem from UCI (University of California Irvine) data repository. We use box plot to compare experimental results. Our proposed method shows better performance in most of the individual experiment.

Key words: C4.5, Entropy, Pearson's Correlation, Spearman Correlation, Kendall Correlation.

1 Introduction

C4.5 is a popular decision tree based algorithm to solve data mining task. Professor Ross Quinlan from University of Sydney has developed C4.5 in 1993 [1]. Basically it is the advance version of ID3 algorithm, which is also proposed by Ross Quinlan in 1986 [2]. C4.5 has additional features such as handling missing values, categorization of continuous attributes, pruning of decision trees, rule derivation and others. C4.5 constructs a very big tree by considering all attribute values and finalizes the decision rule by pruning. It uses a heuristic approach for pruning based on the statistical significance of splits. Basic construction of C4.5 decision tree is [3].

- The root nodes are the top node of the tree. It considers all samples and selects the attributes that are most significant.
- The sample information is passed to subsequent nodes, called 'branch nodes' which eventually terminate in leaf nodes that give decisions.
- Rules are generated by illustrating the path from the root node to leaf node.

Dealing huge data with computational efficiency is one of the major challenges for C4.5 users. Most of the time, it is very difficult to handle data file when dimensionality expands enormously during process for rule generation. As C4.5 uses decision tree, it needs to consider some other issues such as depth of the decision tree, handling of continuous attributes, method of selection measure to adopt significant attributes, dealing of missing values, etc. Following section illustrates about some features of C4.5 algorithm.

1.1 Features of C4.5 Algorithm

There are several features of C4.5. Some features of C4.5 algorithm are discussed below.

1.1.1 Continuous Attributes Categorization

Earlier versions of decision tree algorithms were unable to deal with continuous attributes. 'An attribute must be categorical value' was one of the preconditions for decision trees[3]. Another condition is 'decision nodes of the tree must be categorical' as well. Decision tree of C4.5 algorithm illuminates this problem by partitioning the continuous attribute value into discrete set of intervals which is widely known as 'discretization'. For instance, if a continuous attribute C needs to be processed by C4.5 algorithm, then this algorithm creates a new Boolean attributes C_b so that it is true if $C<b$ and false otherwise [6]. Then it picks values by choosing a best suitable threshold.

1.1.2 Handling Missing Values

Dealing with missing values of attribute is another feature of C4.5 algorithm. There are several ways to

handle missing attributes. Some of these are Case Substitution, Mean Substitution, Hot Deck Imputation, Cold Deck Imputation, Nearest Neighbour Imputation [6]. However C4.5 uses probability values for missing value rather assigning existing most common values of that attribute. This probability values are calculated from the observed frequencies in that instance. For example, let A is a Boolean attribute. If this attribute has six values with A=1 and four with A=0, then in accordance with Probability Theory, the probability of A=1 is 0.6 and the probability of A=0 is 0.4. At this point, the instance is divided into two fractions: the 0.6 fraction of the instances is distributed down the branch for A=1 and the remaining 0.4 fraction is distributed down the other branch of tree. As C4.5 split dataset to training and testing, the above method is applied in both of the datasets. In a sentence we can say that, C4.5 uses most probable classification which is computed by summing the weights of the attributes frequency.

1.2 Limitations of C4.5 Algorithm

Although C4.5 one of the popular algorithms, there are some shortcomings of this algorithm. Some limitations of C4.5 are discussed below.

1.2.1 Empty branches

Constructing tree with meaningful value is one of the crucial steps for rule generation by C4.5 algorithm. In our experiment, we have found many nodes with zero values or close to zero values. These values neither contribute to generate rules nor help to construct any class for classification task. Rather it makes the tree bigger and more complex.

1.2.2 Insignificant branches

Numbers of selected discrete attributes create equal number of potential branches to build a decision tree. But all of them are not significant for classification task. These insignificant branches not only reduce the usability of decision trees but also bring on the problem of over fitting.

1.2.3 Over fitting

Over fitting happens when algorithm model picks up data with uncommon characteristics. This cause many fragmentations is the process distribution. Statistically insignificant nodes with very few samples are known as fragmentations [5]. Generally C4.5 algorithm constructs trees and grows it branches 'just deep enough to perfectly classify the training examples'. This strategy performs well with noise free data. But most of the time this approach over fits the training examples with noisy data. Currently there are two approaches are widely using to bypass this over-fitting in decision tree learning [4]. Those are:

- If tree grows very large, stop it before it reaches maximal point of perfect classification of the training data
- Allow the tree to over-fit the training data then post-prune tree.

However none of those are complete solution of this problem. So in this research we have proposed two tools to reduce the input space of data. The first tool is Entropy of Information Theory and the second is Correlation Coefficient. In this study, we have examined 8 problems from the UCI Repository [7]. The details of the data sets description is provided in Table 1. A Java based machine learning tool Weka3.4 [8] is used to perform the experiment. The machine configuration is Intel Core2 Duo CPU 2.33GHz and 4GB RAM.

The rest of the paper is organized as follows: Section 2 illustrates few recent researches on improvement of C4.5. Section 3 describes briefly about our proposed method and description of tools. Section 4 describes how we design the experiment. Section 5 is about details of data we have used. Section 6 analysis of result and finally we draw conclusions from our research in Section 7.

2 Literature Review

C4.5 is one of the most widely use algorithm for inductive inference because of its efficiency and comprehensive features. As a result, data miners have proposed several techniques for betterment of this algorithm. In this section, we are going to discuss few recent works. Polat and Gunes [14] have offered 'one against all approach' with C4.5. They have conducted experiment with three famous data set namely Dermatology, Image segmentation, Lymphography from UCI. In their experiment they have found excellent accuracy against other algorithms. But did not mention regarding time and the performance against other type of database. In many cases, algorithms are biased by the nature of data files [13]. Jiang and Yu [15] have proposed a hybrid algorithm based on outlier detection and C4.5. They have worked with imbalance data to make them balance using outlier detection then implement C4.5 algorithm. Their proposed algorithm shows good accuracy relatively to other algorithm namely C4.5 and Ripper [16]. But differences of accuracy with other algorithms are not considerably high according to their experiment result. Computational time is not mentioned in this paper as well. Yu and Ai [17] have worked for classification of Remote Sensing (RS) data using rough set and C4.5 algorithm. Their algorithm performs well on that specific data type. Yang [18] has used hierarchical

clustering to limit the decision tree to binary tree to improve traditional C4.5 algorithm. The author's algorithm successfully trim down the number of leaf nodes and improve accuracy. In our proposed improvement of C4.5, we use Entropy and Correlation Coefficients. We use box plot to compare the significance of accuracy and time.

3 Proposed Method

Basic focus of our experiment is to reduce the input space of a data file, roll back the processing time and boost up the percentage of classification accuracy. To do so, we propose popular measurement of Information Theory the Entropy. Entropy finds out the average uncertainty of collection of data. We have used it to find out the central point of the data file. After getting the central point, we have applied the correlation coefficient to choose significant attributes in the data files. Then we have applied C4.5 algorithm on chosen significant attributes. There are brief discussions on the Entropy and three types of correlation coefficient in the following sections.

3.1 Entropy

Information theory (IT) is a widely used topic for computer scientists, cognitive scientists, data miners, statisticians, biologists, and engineers. In information theory, entropy measures the uncertainty among random variables in a data file. Claude E. Shannon [9] has developed the idea of entropy of random variables. He introduced the beginnings of information theory and of the modern age of Ergodic theory. Entropy and related information provides the long term behaviour of random processes that are very useful to analyse data. The behaviour of random process is also a key factor for developing the coding for information theory. Entropy is a measurement of average uncertainty of collection of data when we do not know the outcome of an information source. That means it's a measurement of how much information we do not have. This also indicates the average amount of information we will receive from outcome of an information source. Let X is an attribute, p is each element and j is position of each element of X then calculation for entropy is

$$H(X) = \sum_{j=1}^{k} p_j \log_2 \frac{1}{p_j} \qquad = -\sum_{j=1}^{k} p_j \log_2 p_j \qquad \text{..... (1)}$$

Larger value H(X) indicates that attribute X is more random. On the other hand, attribute with smaller H(X) value implies less random i.e. this attribute is more significant for the data mining. The value of the entropy attains its minimum 0, when all other p_j's are 0. The value reaches its maximum $\log_2 k$, when all p_j's are equal to 1/k.

3.2 Correlation coefficient

Correlation coefficient is one of the major statistical tools to analysis sets of variables and determines their relationships. So that user can make decisions on the basis of provided information by correlation coefficients. Thus it saves millions even billions of dollars for businessman, reduces enormous time for researchers and scale down effort for many other working person in various profession. Researchers have worked on this tool to improve its efficiency by introducing different way of calculation. Among different correlation coefficients, we have chosen three most popular one which are Pearson's, Kendall and Spearman's correlation coefficients. In the following section we have describe briefly about those.

3.2.1 Pearson correlation coefficient

Pearson's correlation coefficient is developed by Karl Pearson [19]. It measures the linear relationship between two variables by comparing their strength and direction. Relationship between two variables is expressed by -1 to +1. If the variables are perfectly linear related by an increasing relationship, the Correlation Coefficient gains the maximum value i.e. +1. On the other hand, if the variables are perfectly linear related by a decreasing relationship, the correlation value gains -1. And a value of 0 expresses that the variables are not linear related by each other. In general, if the correlation coefficient is greater than 0.8, it expresses strong correlation between variables.

Let X and Y are interval or ratio variables. They are normal distribution and their joint distribution is bivariate normal. So the formula of Pearson's Correlation Coefficient is:

$$r_{xy} = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sqrt{(SS_X)(SS_Y)}} \qquad OR$$

$$r_{xy} = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sqrt{\left[\left(\sum X^2 - \frac{(\sum X)^2}{n_x}\right)\left(\sum Y^2 - \frac{(\sum Y)^2}{n_y}\right)\right]}} \qquad \text{..... (2)}$$

Where
ΣX is sum of all the X scores.
ΣY is sum of all the Y scores.

ΣX^2 is square of each X score and then sum of them.
ΣY^2 is square of each Y score and then sum of them.
ΣXY is multiply of each X score by its associated Y score and then add of the resulting products together. This is also called cross product.
n refers to the number of "pairs" of data

3.2.2 Spearman's rank correlation coefficient

Spearman's correlation [20] uses nonparametric method to measure the correlation between variable. It describes the relationship of arbitrary monotonic function of two variables. This correlation does not need frequency distribution of the variables for calculation. Assumption of linear relationship between variable is not required in this correlation. Generally Spearman correlation coefficient is denoted by the Greek letter ρ (rho). It performs well with testing the null hypothesis off the relationship. The range of value of Spearman's correlation coefficient is -1 to +1.

In order to compute the Spearman rank correlation coefficient, the two variables (X and Y) are converted to ranks. A rank is assigned according with the position of value into a sort serried of values. In assignment of rank process, the lowest value had the lowest rank and the highest value has the highest rank. When there are two equal values for two different compounds, the associated rank had equal values and is calculated as means of corresponding ranks. Then we need to calculate the difference between two ranks. Let *d* is the difference of two ranks and *n* is the total pair of variables, the formula of Spearman's correlation coefficient is:

$$\rho = 1 - \frac{6 \times \sum d_i^2}{n(n^2 - 1)} \quad \text{.....(3)}$$

3.2.3 Kendall's rank correlation coefficients

Kendall correlation coefficient [21] is also uses nonparametric method for correlation measure. It is also regarded as Spearman rank correlation coefficient. Spearman correlation is calculated from variables' rank rather Kendall correlation is associated with probability calculation. Kendal Correlation coefficient is denoted with the Greek letter τ (tau). Kendall-tau uses concordant or discordant values. The range of value of Kendall correlation coefficient is -1 to +1.

Let X and Y be the pair of measured and estimated inhibitory activity. Kendall tau coefficient is defined as

$$\tau = \frac{n_c - n_d}{\frac{1}{2} n(n-1)} \quad \text{.....(4)}$$

Where n_c is concordant value, n_d is discordant value and n is total number of instance.

4 Experimental Design

To perform our experiment, we have calculated entropy using Matlab [10] programming tools. We choose the attribute with minimum entropy value. According to entropy property, we nominate that attribute as the central attribute of the database. Then we find out Pearson's, Spearman and Kendall correlation coefficient based on the central attribute using Matlab. Finally we have applied C4.5 algorithm with WEKA [7]. WEKA provides different types of test options to classify data files such as use training set, supplied test set, cross validation and percentage split. We choose 10 fold cross validation if number of instance less than or equal to 1000. In case of more than 1000 instance, we have split data file to 70% training and 30% testing data.

5 Data Description

We have experiment on 8 data files. All these data files are picked up from popular UCI [8] data repository. Table 1. shows the details of those files.

Table 1: Data files properties

Data file Name	Total Instances	Total Attribute *(before improved method applied)*	Total Attribute *(after improved method applied)*
optdigits	5620	65	34
waveFormNoise	5000	41	23
vehicle	846	19	13
ionosphere	351	35	19
Sonar	208	61	33
Glass	214	10	6
wpbc	199	34	21
parkinson	195	23	15

6 Experimental Outcome

Table 2 shows the comparison of modelling time and accuracy among original C4.5, improved Pearson's, improved Spearman and improved Kendall C4.5 algorithm. Improved Pearson's shows the supremacy in modelling time and accuracy for each data file except 'glass'. But improved Spearman C4.5 shows tremendous performance for that specific data file. It is said that Spearman correlation coefficient and Kendall correlation coefficient are similar type of correlation coefficient. However improved Spearman is more consistence than Kendall according to box plot analysis in figure 1.

Table 2 : Comparisons of original C4.5 and three Improved C4.5

Data file Name	C4.5		Improved C4.5 (Pearson's)		Improved C4.5 (Spearman)		Improved C4.5 (Kendall)	
	Modelling Time	*Accuracy*	*Modelling Time*	*Accuracy*	*Modelling Time*	*Accuracy*	*Modelling Time*	*Accuracy*
ionosphere	0.03	80.1887 %	0.02	92.3077 %	0.02	91.453 %	0.02	92.0228 %
waveFormNoise	0.67	84.3601 %	0.41	86.9601 %	0.41	83.9601 %	0.36	83.8201 %
wpbc	0.02	70.3518 %	0.001	74.3719 %	0.001	74.3719 %	0.02	75.8794 %
optdigits	1.27	90.6941 %	0.72	91.3808 %	0.69	92.7367 %	0.72	91.3808 %
vehicle	0.05	72.4586 %	0.02	71.6178 %	0.02	73.2598 %	0.02	71.6178 %
glass	0.22	61.5385 %	0.02	60.000%	0.02	69.2308 %	0.02	60.000%
sonar	0.03	71.1538 %	0.03	71.8347 %	0.03	72.6538 %	0.03	71.8347 %
parkinsons	0.02	80.5128 %	0.02	86.1538 %	0.001	85.6401 %	0.02	84.6154

7 Result Analysis

We have used Box Plot [11], a visual representation of statistical technique with five number analyses to analyse our experimental data. We have applied Matlab [10] to construct the box plot. Figure 1 reflects about comparison among original C4.5 and our improved C4.5 algorithms. According to Box Plot illustration of Figure 1, the median line of box for C4.5 algorithm is at 76%. On the other hand, median line for improved C4.5 with Pearson's, Spearman and Kendall correlation coefficients

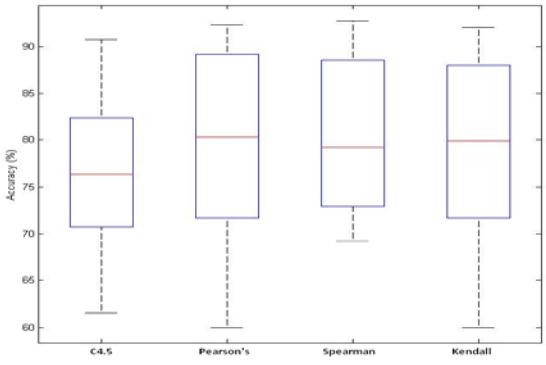

Figure 1: Box plot analysis of accuracy among algorithms

are 81%, 79% and 80% respectively. In regards of dispersion of data, inter-quartile ranges (both upper quartile and lower quartile) are also obtained superior value of box plot. As average performance of all algorithms are good, there are no potential outliers in this graphical chart. However pattern of skewness is not straightforward and not symmetrical for all algorithms. Improved C4.5 with Pearson's correlation coefficient has smaller values with low-skew as it has longer whisker at the bottom of the box. But the box itself is symmetrical which contain the middle 50% of accuracy experimental data of the improved Pearson's correlation coefficient algorithm. This box also obtains highest value of upper quartile among all the algorithms in our experiment. Whiskers of improved C4.5 with Spearman correlation coefficient are symmetrical. Moreover this box appears to be upper-skew, because the line marking of median is towards the bottom of the box. Thus the box indicates that accuracy of this algorithm has more upper values then lower. The box plot reflects that the nature of improved C4.5 with Pearson's and Kendall correlation coefficient are all most similar except a bit long whisker on top of Kendall. On the whole, general C4.5 algorithm has longer whiskers and relatively smaller box in the figure 1which indicates that performance of this algorithm is stagnant within a certain range. Whereas other improved C4.5 algorithms proposed in this paper are significantly better than the original C4.5 algorithm.

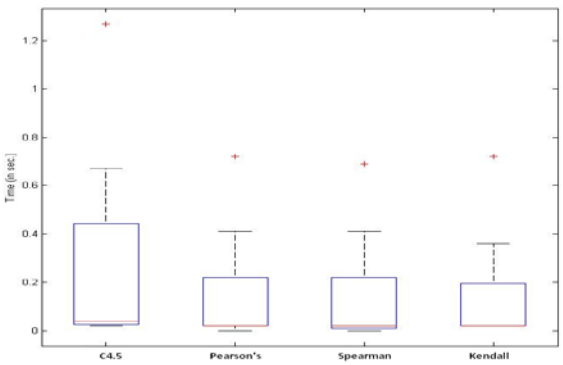

Figure 2: Box plot analysis of processing time among algorithms

Figure 2 reveals comparison of processing time among C4.5 and improved C4.5 algorithms. At a glance we can explicate that our proposed C4.5 algorithms takes less processing time than original C4.5. There is an outlier for each box in the plot because of relatively large data file. Original C4.5 has the highest value (1.27 sec) than other improved C4.5 algorithms. However, nowadays high performance computer, super computer, etc. are available for users. which lessen processing timing tremendously.

8 Conclusion

In this research, we have proposed to improve a rule-base classification algorithm C4.5. The main objective of this research is to boost up the classification accuracy and simultaneously roll back timing to build a classification model. We have emphasized reducing reduce input space using entropy and several correlation coefficients formulas. The proposed method shows better performance for each data file. However, individually each improved C4.5 is not performing better than original C4.5 in every test case. Improved Pearson's C4.5 is most consistent among three. Between improved Spearman C4.5 and improved Kendall C4.5, Spearman shows the better performance in our experiment. We aim to continue this research by analysing the data file we have investigated. We will find out why it is performing better in one proposed method but not performing well on other one. We will consider more data file to get a better outcome of this experiment.

References:

[1] J.R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, San Mateo, CA, 1993.

[2] J.R.Quinlan, Induction of Decision Trees, *Machine Learning*, 1986, pp81-106.

[3] A. B. M. S. Ali and S. A. Wasimi, Data Mining: Methods and Techniques, Thomson Publishers, Victoria, Australia, 2007.

[4] A. B. M. S. Ali and K. A. Smith, On learning algorithm for classification, *Applied Soft Computing*, Dec 2004. pp. 119-138.

[5] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publish, 2001.

[6] M. Singh, How to Handle Missing Values, Articlebase, viewed on Oct 2009, at http://www.articlesbase.com/information-technology-articles/how-to-handle-missing-values-538449.html#.

[7] C. Blake and C.J. Merz, UCI Repository of machine learning databases, University of California Irvine, 2007. - Feb 2008. - http://archive.ics.uci.edu/ml/.

[8] I.H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tool and Technique with Java Implementation*, San Francisco: Morgan Kaufmann, 2000.

[9] C. E. Shannon, A Mathematical Theory of Communication, *The Bell System Technical Journal*, 30:50-64, January 1948.

[10] Matlab, *Statistics Toolbox User's Guide*, The MathWorksInc, USA . 2008. Version 6.2

[11] J. W. Tukey, *Exploratory Data Analysis*, Addison-Wesley Publishing Company, 1977.

[12] M.M. Mazid, A.B.M. S.Ali, , and K.S. Tickle, A Comparison Between Rule Based and Association Rule Mining Algorithms, *In Proceedings of the IDSS-NDS conference*, Gold Coast, Australia, Oct. 2009.

[13] M. M. Mazid, S. Ali, and K.S. Tickle, 2008. Finding an unique Association Rule Mining Algorithm based on data characteristics, *In Proceedings of the IEEE/ICECE*, Dec 2008, Dhaka, Bangladesh.

[14] K. Polat and S. Güne, A novel hybrid intelligent method based on C4. 5 decision tree classifier and one-against-all approach for multi-class classification problems, *Expert Systems with Applications*, vol. 36, 2009, pp. 1587-1592

[15] S. Jiang and W. Yu, A Combination Classification Algorithm Based on Outlier Detection and C4. 5, *Springer Publications*, 2009.

[16] W. W. Cohen, Fast effective rule induction, *In Proceedings of the Twelfth International Conference on Machine Learning Chambery*, France., 1993, pp. 115–123.

[17] M. Yu and T. H. Ai, Study of RS data classification based on rough sets and C4. 5 algorithm, *In Proceedings of the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series* , 2009.

[18] X. Y. Yang, Decision tree induction with constrained number of leaf node, Masters Thesis, National Central University, Taiwan, 2009.

[19] K. Pearson, Notes on the history of correlation, *Biometrika*, 1920, vol. 13, pp. 25-45.

[20] C. Spearman, The proof and measurement of association between two things, *The American journal of psychology*, 1904, pp. 72-101.

[21] M.G. *Kendall Rank Correlation Methods*, Hafner Publishing Co, New York, 1955.