# IMPROVED C4.5 DECISION TREE CLASSIFIER ALGORITHM FOR ANALYSIS OF DATA MINING APPLICATION

**[1]Ms. Gayatri V. Badgujar, [2]Prof. Khushboo Sawant**

**[1]PG Student, [2]Assistant Professor, [1,2]Department of Computer Science and Engineering, Jagadguru Dattetrya College of Technology, Indore, Madhya pradesh, India.**

Abstract: **Decision tree is an important method for both induction research and data mining, which is mainly used for model classification and prediction. ID3 and C4.5 algorithm is the most widely used algorithm in the decision tree .Illustrating the basic ideas of decision tree in data mining, in this paper ,shortcomings of ID3's and C4.5 inclining to choose attributes with many values is discussed , and then a new decision tree algorithm presented .Experimental results show that the proposed algorithm can overcome ID3's and C4.5 shortcoming effectively and get more reasonable and effective rules. C4.5 is one of the most classic classification algorithms on data mining, but when it is used in mass calculations, the efficiency is very low. In this paper, the rule of C4.5 is improved by the use of L'Hospital Rule, which simplifies the calculation process and improves the efficiency of decision making algorithm. We aim to implement the algorithms in a very time and space effective manner and throughput and response time for the application will be promoted as the performance measures. Our project aims to implement these algorithms and graphically compare the complexities and efficiencies of these algorithms.**

***Keywords: Data mining, decision tree,ID3,algorithm C4.5,L'Hospital Rule ,the rate of information gain ,large datasets.***

## I. INTRODUCTION

With the development of computer technology and computer network technology, the degree of information is getting higher and higher, people's ability of using information technology to collect and produce data is substantially enhanced. How can we not be drowned by the sea of information, and from which discovering useful knowledge and improving the effectiveness of information utilization are problems need to be addressed urgently. Data mining is a process to extract information and knowledge from a large number of incomplete, noisy, fuzzy and random data. In these data, the information and knowledge are implicit, which people do not know in advance, but potentially useful. At present, the decision tree has become an important data mining method. The basic learning approach of decision tree is greedy algorithm, which use the recursive top-down approach of decision tree structure. Quin lan in 1979 put forward a well-known ID3 [1],[ 2], [3] algorithm, which is the most widely used algorithm in decision tree. But that algorithm has a defect of tending to use attributes with many values. Aiming at the shortcomings of the ID3 algorithm, in the paper, we analyzed several decision tree classification algorithms currently in use, including the ID3 [4] and C4.5 [2] algorithm as well as some of the improved algorithms [3], [5] ,[6] thereafter them. When these classification algorithms are used in the data processing, we can find that its efficiency is very low and it can cause excessive consumption of memory. On this basis, combining with large quantity of data, we put forward the improvement of C4.5 algorithm efficiency, and uses L'Hospital rule to simplify the calculation process by using approximate method. This improved algorithm not only has no essential impact on the outcome of decision-making, but can greatly improve the efficiency and reduce the use of memory. So it is more easily used to process large amount of data collection.

### *1.1 Decision Tree*

Decision trees are built of nodes, branches and leaves that indicate the variables, conditions, and outcomes, respectively. The most predictive variable is placed at the top node of the tree. The operation of decision trees is based on the ID3 or C4.5 algorithms. The algorithms make the clusters at the node gradually purer by progressively reducing disorder (impurity) in the original data set. Disorder and impurity can be measured by the well-established measures of entropy and information gain. One of the most significant advantages of decision trees is the fact that knowledge can be extracted and represented in the form of classification (if-then) rules. Each rule represents a unique path from the root to each leaf. In operations research, specifically in decision analysis, a decision tree (or tree diagram) is a decision support tool that uses a graph or model of decisions and their possible consequences. A decision tree is used to identify the strategy most likely to reach a goal. Another use of trees is as a descriptive means for calculating conditional probabilities. A decision tree is a flow-chart-like tree structure, where each branches represents an out-come of the test, and each leaf node represent classes. The attribute with highest information gain is chosen as test attribute for current node. This attribute minimizes the information needed to classify the sample. A decision tree is a tree in which each branch node it represents a choice between a number of alternatives, and each leaf node represents a decision. Decision tree are commonly used for gaining information for the purpose of decision -making. Decision tree starts with a root node on which it is for users to take actions. From this node, users split each node recursively according to decision tree learning algorithm. The final result is a decision tree in which each branch represents a possible scenario of decision and its outcome.

## II. METHODOLOGY

### 2.1 Modules of the System

**Select Dataset:** Selecting a dataset actually includes giving the dataset as an input to the algorithm for processing.

**ID3 Processing:** ID3 processing includes the processing the given input dataset according to the defined algorithm of ID3 data mining.

**C4.5 Processing:** C4.5 processing includes the processing the given input dataset according to the defined algorithm of C4.5 data mining.

**Improved C4.5 Processing:** Improved C4.5 processing includes the processing the given input dataset according to the defined algorithm of improved C4.5 data mining.

**Generate Trees:** The data which should be inputted to the tree generation mechanism is given by the ID3, C4.5 and improved C4.5 processors. Tree generator generates the tree for ID3, C4.5 and improved C4.5 decision tree algorithm.

### 2.2 Attribute Selection Measure

The information gain measure is used to select the test attribute at each node in the tree. The attribute with highest information gain is chosen as test attribute for the current Node. This attribute minimizes the information needed to classify the samples in resulting partition and reflect the least “impurity” in these partitions.

Let S be set consisting of data sample. Suppose the class label attribute has m Distinct values defining m distinct class $C_i$ (for i =1... m).

Let $S_i$ be the number of Sample of S in class $C_i$. The expected information needed to classify a given sample is given by equation

$$I\ (S1,\ S2,\ \cdot\cdot\cdot\ ,\ Sm) = -\sum_{i=1}^{m} P_i \log_2(P_i)$$

Where Pi is probability that an arbitrary sample belongs to classify Ci and estimated by $S_i$/S. Note that a log function to base 2 is used since the information in encoded in bits[10] .

### 2.3 Entropy

It is minimum number of bits of information needed to encode the classification of arbitrary members of S. Lets attribute A have v distinct value $a_1$,............., $a_v$. Attribute A can be used to Partition S into v subsets, $S_1$, $S_2$,........, $S_v$ , where $S_j$ contains those samples in S that have value $a_j$ of A. If A were selected as the test attribute, then these subset would corresponds to the branches grown from the node contains the set S. Let $S_{ij}$ be the number of class $C_i$, in a subset by $S_{j}$ , The entropy or expected information based on partitioning into subset by A, is given by equation

$$E(A) = \sum_{j=1}^{v} (S_{1j} + S_{2j} + \cdots + S_{mj} / S)*I(S_{ij} + \cdots + S_{mj}) .$$

The first term acts as the weight of the jth subset and is the number of samples in the subset divided by the total number of sample in S. The smaller the entropy value, the greater purity of subset partitions as shown in

$$I(S_1, S_2, \cdots, S_m) = -\sum_{i=1}^{m} P_i \log_2(P_i)$$

Where $P_i$ is the probability that a sample in $S_j$ belongs to class $C_i$[10].

### 2.4 Information Gain

It is simply the expected reduction in entropy caused by partitioning the examples according to the attribute .More precisely the information gain, Gain(S, A) of an attribute A, relative collection of examples S, is given by equation.

$$\text{Gain}(A) = I(S_1, S_2, \cdots, S_m) - E(A)$$

In other words gain (A) is the expected reduction in entropy caused by knowing the Value of attribute A. The algorithm computes the information gain of each attribute. With highest information gain is chosen as the test attribute for a given set[10].

**The strengths of decision tree methods:**

1. Decision trees are able to generate understandable rules.
2. Decision trees perform classification without requiring much computation.
3. Decision trees are able to handle both continuous and categorical variables.
4. Decision trees provide a clear indication of which fields are most important for prediction or classification.

## III. ID3 ALGORITHM

ID3 is a simple decision tree learning algorithm developed by Ross Quinlan (1983). The basic idea of ID3 algorithm is to construct the decision tree by employing a top-down, greedy search through the given sets to test each attribute at every tree node. In order to select the attribute that is most useful for classifying a given sets, we introduce a metric information gain.

### 3.1 The shortcoming of ID3 algorithm

The principle of selecting attribute A as test attribute for ID3 is to make E (A) of attribute A, the smallest. Root node is decided only on value of information gain of attribute .Missing values of the attribute is not considered in ID3 algorithm .It is not so important in real situation for those attributes selected by ID3 algorithm to be judged firstly according to make value of entropy minimal. Besides, ID3 algorithm selects attributes in terms of information entropy which is computed based on probabilities, while probability method is only suitable for solving stochastic problems[11].

## IV. C4.5 ALGORITHM

Quinlan puts forward ID3 decision tree algorithm base on the information gain, and later, an improved C4.5 algorithm in1993. Many scholars made kinds of improvements on the decision tree algorithm. But the problem is that these decision tree algorithms need multiple scanning and sorting of data collection several times in the construction process of the decision tree. The processing speed reduced greatly in the case that the data set is so large that can not fit in the memory. At present, the literature about the improvement on the efficiency of decision tree classification algorithm For example, *Wei Zhao, Jamming Su* in the literature [7] proposed improvements to the ID3 algorithm, which is simplify the information gain in the use of Taylor's formula. But this improvement is more suitable for a small amount of data, so it's not particularly effective in large data sets.

Due to dealing with large amount of datasets, a variety of decision tree classification algorithm has been considered.

The advantages of C4.5 algorithm is significantly, so it can be choose. But its efficiency must be improved to meet the dramatic increase in the demand for large amount of data.

## V. THE IMPROVEMENT OF C4.5 ALGORITHM

### *5.1 The improvement*

The C4.5 algorithm [8] [9] generates a decision tree through learning from a training set, in which each example is structured in terms of attribute-value pair. The current attribute node is one which has the maximum rate of information gain which has been calculated, and the root node of the decision tree is obtained in this way. Having studied carefully, we find that for each node in the selection of test attributes there are logarithmic calculations, and in each time these calculations have been performed previously too. The efficiency of decision tree generation can be impacted when the dataset is large. We find that the all antilogarithm in logarithmic calculation is usually small after studying the calculation process carefully, so the process can be simplified by using L'Hospital Rule. As follows:

If f(x) and g(x) satisfy:

(1) $\lim_{x\to x0} f(x)$ And $\lim_{x\to x0} g(x)$ are both zero or are both ∞

(2) In the deleted neighbourhood of the point x0, both f'(x) and g'(x) exist and g'(x)! = 0;

(3) $\lim_{x\to x_0}\frac{f(x)}{g(x)}$ Exist or is ∞ then

$\lim_{x\to x_0}\frac{f(x)}{g(x)} = \lim_{x\to x_0}\frac{f'(x)}{g'(x)}$ so

$$\lim\frac{[\ln(l-x)]'}{-x} = \lim\frac{\ln(l-x)}{-x} = \lim\frac{-\frac{l}{l-x}}{-l} = \lim\frac{l}{l-x} = 1$$

X approaches 0 viz. $\ln(1-x) = -x$ (x approaches 0)

(1) $\ln(1-x) \approx -x$ (when x is quite small)

(2) Suppose c = 2, that is there are only two categories in the basic definition of C4.5 algorithm. Each candidate attribute's information gain is calculated and the one has the largest information gain is selected as the root. Suppose that in the sample set S the number of positive is p and the negative is n. So we can get the equation :

$$\text{E(S,A)}= \sum_{j=1}^{r}\frac{p_j - n_j}{p+n} I(S_{1j} + S_{2j})$$

in which $p_j$ and $n_j$ are respective the number of positive examples and the negative examples in the sample set .so gain Ratio(A) can be simplified as :

$$Gain-Ratio(A) = \frac{Gain(A)}{I(A)} = \frac{E(S)-E(S,A)}{I(A)}$$

$$\frac{I(p,n)-\{\frac{S_1}{N}I(S_{11},S_{12})+\frac{S_2}{N}I(S_{21},S_{22})\}}{I(S_1,S_2)}$$

S1: the number of positive examples in A

S2: the number of negative examples in A

S11: the number of examples that A is positive and attributes value is positive,

S12: the number of examples that A is positive and attributes value is negative,

S21: the number of examples that A is negative and attributes value is positive,

S22: the number of examples that A is negative and attributes value is negative

Go on the simplification we can get:

$$\{\frac{P}{N}\log_2\frac{P}{N}+\frac{n}{N}\log_2\frac{n}{N}-\{\frac{S_1}{N}[\frac{S_{11}}{S_1}\log_2\frac{S_{11}}{S_1}+$$

$$\frac{S_{12}}{S_1}\log_2\frac{S_{12}}{S_1}]+\frac{S_2}{N}[\frac{S_{21}}{S_2}\log_2\frac{S_{21}}{S_2}+\frac{S_{22}}{S_2}\log_2\frac{S_{22}}{S_2}]\}$$

$$/\{\frac{S_1}{N}\log_2\frac{S_1}{N}+\frac{S_2}{N}\log_2\frac{S_2}{N}\}$$

In the equation above, we can easily learn that each item in both numerator and denominator has logarithmic calculation and N, Divide the numerator and denominator by log2e simultaneously, and multiplied by N simultaneously. we can get equation:

Gain-Ratio(S,A)

$$=\{p\ln\frac{P}{N}+n\ln\frac{n}{N}-\{[S_{11}\ln\frac{S_{11}}{S_1}+S_{12}\ln\frac{S_{12}}{S_1}]$$

$$+[S_{21}\ln\frac{S_{21}}{S_2}+S_{22}\ln\frac{S_{22}}{S_2}]\}/\{S_1\ln\frac{S_1}{N}+S_2\ln\frac{S_2}{N}\}$$

Because N= p + n, $\frac{P}{N}+\frac{n}{N}=1.$

Then replaces $\frac{P}{N}$ and $\frac{n}{N}$ with $1-\frac{n}{N}$ and $1-\frac{P}{N}$ respectively, then we can get equation:

$$Gain-Ratio(S,A)=\{P\ln(1-\frac{P}{N})-\{[S_{11}\ln(1-\frac{S_{12}}{S_1})+$$

$$S_{12}\ln(1-\frac{S_{11}}{S_1})]+[S_{12}\ln(1-\frac{S_{22}}{S_2})+S_{22}\ln(1-\frac{S_{21}}{S_2})]\}$$

$$/S_1\ln(1-\frac{S_2}{N})+S_2\ln(1-\frac{S_1}{N})$$

Because we already have equation (2) so we get ; Gain-Ratio(S,A)= $\dfrac{\frac{Pn}{N}-\{[\frac{S_{11}*S_{12}}{S_1}]+[\frac{S_{21}*S_{22}}{S_2}]\}}{\frac{S_1S_2}{N}}$

In the expression above, Gain-Ratio (A) only has addition, subtraction, multiplication and division but no logarithmic calculation, so computing time is much shorter than the original expression. What's more, the simplification can be extended for multi-class.

### *5.2 Reasonable arguments for the improvement*

In the improvement of C4.5 above, there is no item increased or decreased only approximate calculation is used when we calculate the information gain rate. And the antilogarithm in logarithmic calculation is a probability which is less than 1. In order to facilitate the improvement of the calculation, there are only two categories in this article and the probability is a little bigger than in multi-class. And the probability will become smaller when the number of categories becomes larger; it is more helpful to justify the rationality. Furthermore, there is also the guarantee of L'Hospital Rule in the approximate calculation, so the improvement is reasonable.

### *5.3 Comparison of the complexity*

To calculate Gain – Ratio(S, A), the C4.5 algorithm's complexity is mainly concentrated in E(S) and E(S, A).When we compute E(s), each probability value is needed to calculated first and this need o (n) time. Then each one is multiplied and accumulated which need O($\log_2$n) time. So the complexity is O($\log_2$n).Again, in the calculation of E(S,A),the complexity is O(n($\log_2$n)$^2$),so the total complexity of Gain-Ration(S,A) is O(n($\log_2$n)$^2$).And the improved C4.5 algorithm only involves original data and only addition, subtract, multiply and divide operation. So it only needs one scan to obtain the total value and then do some simple calculations, the total complexity are O (n)[12].

## VI. EXPERIMENTAL RESULTS

Experimental data is collected from UCI machine learning repository ,which is publicly available .The results were analysis using ID3 ,C4.5 and Improved C4.5 decision tree algorithm to test accuracy and time complexity of classifiers .To observe the performance of the classifiers on large datasets in terms of node count and rule count and time complexity are presented in the Table 1, Table 2 and Table 3 .The result of experiment shows that the effect of improved C4.5 is better than the ID3 ,C4.5 in three aspects such as node count rule count and time complexity .Time is saved because its complexity is changed from O(n($\log_2$n)$^2$) to O(n). and also the improved C4.5 does not need to scan the for several times ,the memory is also saved .for Showing the change clearly we transform the Table 2, Table 3 and Table 1 to the graph in figure 1 and figure 2 [13].

**Table 1. Time complexity Comparison of Id3,C4.5 and Improved C4.5**

| Dataset | Record Num n | Time(msec.) | | |
|---|---|---|---|---|
| | | ID3 | C4.5 | Improved C4.5 |
| Db1 | 170 | 30 | 107 | 66 |
| MyDataSet | 768 | 50 | 1125 | 322 |
| Sales | 10 | 15 | 31 | 19 |
| MyVote | 233 | 42 | 1035 | 99 |

**Table 2. Comparison of ID3,C4.5 and Improved C4.5**

| Dataset | Record Num n | Rules(count) | | |
|---|---|---|---|---|
| | | ID3 | C4.5 | Improved C4.5 |
| Db1 | 170 | 8 | 13 | 8 |
| MyDataSet | 768 | 95 | 114 | 86 |
| Sales | 10 | 10 | 13 | 10 |
| MyVote | 233 | 30 | 93 | 30 |

**Table 3. Comparison of ID3,C4.5 and Improved C4.5**

| Dataset | Record Num n | node(count) | | |
|---|---|---|---|---|
| | | ID3 | C4.5 | Improved C4.5 |
| Db1 | 170 | 4 | 8 | 4 |
| MyDataSet | 768 | 48 | 62 | 45 |
| Sales | 10 | 3 | 5 | 3 |
| MyVote | 233 | 15 | 39 | 16 |

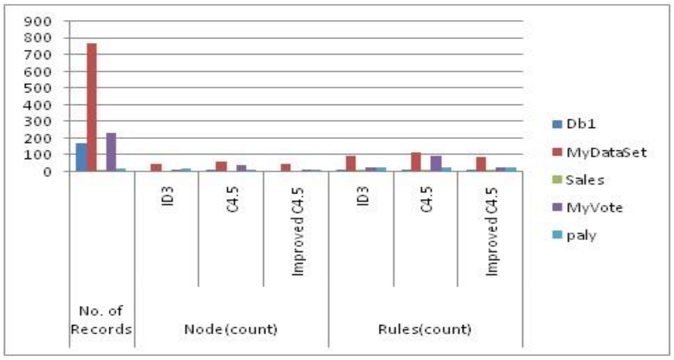

**Fig.1 Comparison of ID3, C4.5 and Improved C4.5**

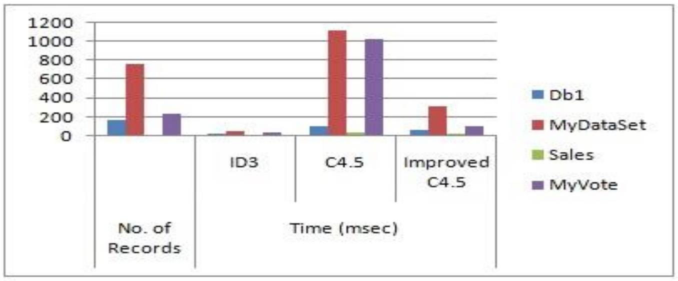

**Fig.2 Comparison of time complexity of ID3,C4.5 and Improved C4.5**

## VII. CONCLUSION

In this Paper C4.5 algorithm was improved and we use approximate calculation of Gain-Ratio(S,A) the experiment proved that it has minimal impact on the classification accuracy ,but the efficiency increased a lot .We can not only speed up the growing of the decision tree , so that better information of rules can be generated. In this paper the algorithm was verified by different large datasets which are publicly available on UCI machine learning repository. With the improved algorithm ,we can get faster and more effective results without the change of the final decision and the presented algorithm constructs the decision tree more clear and understandable .Efficiency and classification is greatly improved and the disadvantages of low efficiency and memory consumption while dealing with large amount of data were overcome as it was in C4.5.If the amount of data is small the original C4.5 is used because of its higher accuracy.

## REFERENCES

[1] I. H. Witten, E. Frank, Data Mining Practical Machine Learning Tools and Techniques, China m/c Press, 2006.

[2] S. F. Chen, Z. Q. Chen, Artificial intelligence in knowledge engineering [M]. Nanjing University Press, 1997.

[3] Z. Z. Shi, Senior Artificial Intelligence [M]. Beijing: Science Press,1998.

[4] D. Jiang, Information Theory and Coding [M]: Science and Technology of China University Press, 2001.

[5] M. Zhu, Data Mining [M]. Hefei: China University of Science and Technology Press ,2002.67-72.

[6] A. P. Engelbrecht., A new pruning heuristic based on variance analysis of sensitivity information[J]. IEEE Trans on Neural Networks, 2001, 12(6): 1386-1399.

[7] N. Kwad, C. H. Choi, Input feature selection for classification problem [J],IEEE Trans on Neural Networks, Quinlan JR. Induction of decision tree [J].Machine Learing.1986

[8] UCI Repository of machine earning databases. University of California, Department of Information and Computer Science, 1998. http: //www.ics. uci. edu/～mlearn/MLRepository. Html

[9] UCI Machine Learning Repository –http://mlearn.icsuci.edu/ database

[10] Jaiwei Han and Micheline Kamber , Data Mining Concepts and Techniques.Morgan Kaufmann Publishers.

[11] Chen Jin,Luo De –lin,mu Fen-xiang ,An Improved ID3 Decision tree algorithm,Xiamen University,2009.

[12] Rong Cao, Xu,Improved C4.5 Decision tree algorithm for the analysis of sales.Southeast University Nanjing,china,2009.

[13] Huang Ming, NiuWenying ,Liang Xu ,An improved decision tree classification algorithm based on ID3 and the application in score analysis.Dalian jiao Tong University,2009.