Modified C4.5 Algorithm with Improved Information Entropy and Gain Ratio

[1] *S.Santhosh Kumar, Research Scholar, Prist University, Thanjavur, (Lecturer, Government College for women, Kumbakonam), Tamil Nadu, India,*
[2] *Dr.E.Ramaraj, Director, Computer Center, Alagappa University, Karaikudi, Tamil Nadu – India,*

Abstract

C 4.5 is the best known classification algorithm used to generate decision trees for continuous and discrete attributes. Although its extension C5.0 is further developed, it is still used in variety of applications. It works on information gain ratio of given attributes. The limitations of C4.5 is its information entropy, it gives poor results for larger distinct attributes. In this work we proposed a modified approach to overcome the draw backs of C4.5 based on information gain and gain ratio.

Key words: - C4.5 Algorithm, Information gain, Entropy, Gain Ratio.

1. Introduction

Decision Trees (DT) are learning algorithms, work based on processing and deciding upon attributes of the data. Attributes in DT are nodes and each leaf node is representing a classification. Decision tree algorithms begin with a set of cases in which each case consists of set of attributes. The attributes are categorized based on its features. The features are represented in the form of symbols or numerical values. All sub sets of attributes are structured on training case. For tree construction, each case (test data) is associated with its training case. The tree formation is based decision making of association of test data with training data. Generally test data has unique feature named as "Non - labelled" data and training data has named as "labelled data". The C4.5 algorithm is a univariant Decision Tree algorithm used decision tree generation.

2. C4.5 Algorithm

The C4.5 is an extension of ID3 which is a similar tree generation algorithm. The basic strategy in ID3 is to selection of splitting attributes with the highest information gain first. That is the amount of information associated with an attribute value that is related to the probability of occurrence. Once the attributes have chosen then amount of information is measured, which is known as entropy. Entropy is used to measure the amount of uncertainty, surprise, or randomness in a dataset. The entropy will be zero if when all of the data in the set belongs to a single class.

3. Related Studies

The study of information theory begins in the year 1924,Harry Nyquist, a researcher at Bell Laboratories, published a work entitled Certain Factors Affecting Telegraph Speed; stated that communication channels had maximum data transmission rates, and he derived a formula for calculating these rates in finite bandwidth noiseless channels. Another remarkable research was done by Nyquist's friend R.V.L.Hartley in 1928; in the paper entitled Transmission of Information established the first mathematical foundations for information theory. The real origin of modern information theory is developed in 1948 by Claude Elwood Shannon [1] founded an information theory; it is used to calculate how much information is available in an event. He is known as father of information theory [1]. His contribution is a landmark that one can find the information based on its entropy. Entropy or Shannon entropy is an expected value for needed information. In the same year based on Shannon's Information theory, Norbert Wiener [3] proposed "A Mathematical Theory of Communication" which stated to encode information in communication. In the year 1949, Shannon proposed a another paper in communication theory of secrecy systems for transferring information in secured manner. In 1951, he developed an entropy based language analysis tool called Prediction and Entropy of Printed English. In 1977, G.J. ChaitinIn proposed algorithmic information theory, which is applied in information-theoretic and probabilistic ideas to recursive function theory. In 1986 [2], John Ross Quinlan a computer science researcher developed a new algorithm called ID3 (Iterative Dichotomiser3), used to generate decision trees. He used two basic concepts for finding relationships among the data, namely information gain and entropy. The drawbacks of ID3 such as over fitting, time latency and complex to handle continuous attributes are resolved as new algorithm called C 4.5 by John Ross Quinlan in 1993 [4]. The C

4.5 overcomes the drawbacks of ID3 in many ways. The information theory is the base for many algorithms in various fields especially its contribution is highly in machine learning and data mining applications. . In this work we proposed a new technique to improve the information gain and gain ratio in C 4.5 algorithm.

4. Information Theory

It is a mathematical representation with conditions and parameters that affects the transmission and processing of information. Information theory is based on probability theory and statistics. The important quantities of information are entropy. The Shannon denoted the information entropy as *H* (Eta).The estimation of expected information (entropy) needed to classify a tuple in D:

$$I(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

Here Pi is the probability that an arbitrary tuple in D belongs to class Ci, It is estimated by |Ci, D|/|D|

4.1. Information Gain Ratio

The information gain is good when it is used of small or medium number of values. In the case of large values it returns less information. The gain value is failed to achieve overall gain value. To overcome the limitations C 4.5 uses Gain ratio by splitting the training sets based on its test attributes. The gain ratio takes number and size of branches into account when choosing an attribute. It corrects the information gain by taking the intrinsic information of a split into account. The Gain Ratio can be defined as

$$Gain(A) = Info(D) - Info_A(D)$$

The information gain is defined based on the splitting criterion. The Split Information can be defined as

$$SplitInfo_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$$

The split information is potential information generated by splitting the training data set D, It gives positive results to most of the applications but it has strong limitation when it is used in decision trees. Although information gain is usually a good measure for deciding the relevance of an attribute, but the result is not guaranteed for all kinds of attributes. According to Ross Quinlan's C 4.5 Decision tree classifier; it can able to handle both discrete and continuous attributes. For example consider a consumer data base, a splitting can be possible with highest information gain ratio, attributes such as age, income, repayment capacity etc., The lower gain ratio attributes such as previous loan clearance, family income, family size could not be considered. Suppose a customer have good repayment track of his previous loan is considered to be added weightage for new loan approval. For biological data sets, some of the attributes are always hold best information gain and gain ratio (symptoms, cause etc) but every patient has individual additional information which helps to predict the root cause and remedy for that disease. The adding of additional information with training data can be used to improve the features of splitting criterion which gives better results.

4.2. Limitations of C 4.5 with Information Entropy

Based on the study of information gain and gain ration, the remarks of information entropy in C 4.5 is discussed below

1) It gives very poor results when large distinct values are used in both continuous and discrete attributes.
2) There is no specific measurement technique available to predict actual information gain before it is applied. The information gain ratio is evaluated only from after generation of attribute values. The mismatch or wrong selection of attributes may give less performance and accuracy rate.
3) When the information gain is less than amount of attributes used, then it becomes failure
4) One of the important issues in decision tree information entropy is uncertainty. If the previously chosen attribute has less value; then it is more complex to choose the next. It leads to unconditional selection of attributes.
5) When same valued attributes are used in decision tree generation, the split up is an complex task; gives unbalanced trees.

5. Proposed Methodology

The Gain Ratio needs overcompensate. Sometimes it creates a dependency to choose an attribute depends on its splitting criterion. The important drawback is, it considers only the attribute which is greater than the average information gain. This phenomenon fails to consider the attributes which are having less information gain but it is more important for splitting. That means it gives additional information that may be basic information used to split the attributes accurately. The recommendations to overcome the limitations in Gain Ratio are

1) Addition of numeric attributes, that could be used as additional information for splitting.
2) Accept the presence of anomaly data with minimum error to increase the gain ratio. From the above recommendations, we propose a certain constraints for splitting

Let v denotes the number of splitting based on the training sets. For solving the above said problems. We modified the splitting function as

$$SplitInfo_A(D) = \sum_{j=1}^{v \geq x} \frac{|D_j|}{|D|} \times Info(D_j)$$

Where

$$x = I_A(D) = |v|$$

$$\{\text{only if A satisfies I (D) with v partitions}\}$$

Otherwise

$$x = I_A(D) = |(I_{A1}(D) + I_{A2}(D))|$$

$$\ldots\ldots\ldots + |I_{An}(D))| = |v|$$

$$\{\text{only if } A \geq I(D)\}$$

Hence x is constraint, satisfies only if v attains high information gain with single attribute; otherwise the additional training attributes must be added to achieve high information gain. Here |v| is a non negative integer used for splitting. The general information gain calculated

$$Gain(A) = Info(D) - Info_A(D)$$

The selection of training attribute with supporting attribute leads to improve the performance of C 4.5 algorithm. Based on that we proposed C 4.5 algorithm with modifications called M C4.5 Algorithm.

5.1.The M C 4.5 Algorithm

Algorithm Generate Modified C 4.5 decision tree

Generate a decision tree with highest information gain from Data D.

Input

D is the partition of data set containing training tuples with set of attributes

The procedure used of splitting criterion that highest information gain of A with single or multiple values. ***A*** *is splitting criterion with information gain*

Output: *Decision tree with high information gain*

Method

Create a node N;

If attributes in D are all in the same category called C then

Return and declare N as leaf node with label C;

Else if N is empty then

Return N is a leaf node label with D;

Apply the attribute selection method (D, attribute_list) to find the best splitting_criterion;

If training set is partitioned with single attribute A

Find the gain ratio for A with respect to training sets;

Return and assign as best splitting criterion

Label node N with Splitting criterion

Else add more attribute in the training set to find best splitting_criterion

Find the gain ratio for A with respect to training sets;

Label node N with Splitting_criterion

If splitting criterion is distribute_valued and

Mutli_splits allowed then

attribute_list ← splitting attribute; //splitting attribute value is reduced in overall information gain

For each set of splitting attributes called j as splitting criterion

Let Dj be the set data with set of attributes in D satisfying outcome j;

If Dj is empty then

Attach a leaf label with majority of class in D to node N;

Else

Attach the node returned by Generate_decisioning_tree (Dj, attribute_list) to node N;

End for

Return N;

5.2.Proof by Example

In this section we take a data set to implement our new proposed methodology of splitting criterion gives better results for M C4.5. The dataset contains drug intake person records of different age. The record contains the mixed values of age, year and their Hispanic origins.

From the example, let us consider training set as D, which contains mixed attributes. The class label sex, white, American, Mexican are distinct values.

Table (1).Data set contains Drug Consumption in American Cities

Source: CDC, Atlanta, USA

The possible values based on probability gain are {male, female, both, under age, adult, middle age, and old} with different ranges. It is hard to compute the above class attributes, for that we take disease cause as discrete class values namely {yes, no, NA}. Note that NA is class which does not contains any information about disease cause. so that class must be excluded in the case of use of C 4.5 classification algorithm. Therefore the M value = 2 and class C=

Sex	White	American	Mexican	Disease
Male	34.2	31.1	27.5	yes
Female	47.6	41.4	36	yes
male	41.4	31.2	24	NA
Both (up to 18 years)	22.9	14.8	16.1	No
Adult (18 to 44 years)	34.3	27.8	21.1	yes
Middle Age (45 to 64 years)	55.5	57.5	48.1	yes
Old (65 years and over)	74	74.5	67.7	NA

C_1, C_2, respectively. The expected information can be calculated as

$$I(D) = -\frac{4}{7}\log_2(\frac{4}{7}) - \frac{1}{7}\log_2(\frac{1}{7})$$

$$I(D) = 0.888bits$$

The expected information needed to find tuples of D with respect to male is

$$I(D) = -\frac{3}{7}\log_2(\frac{3}{7})$$

$$I(D) = 0.523bits$$

The expected information needed to find tuples of D with respect to male is

$$G(male) = I(D) - \mathrm{I}_{male}(D)$$
$$= 0.888 - 0.523 = 0.365bits$$

From above example the I(D) is calculated with three classes namely { yes, no, N/A }, here NA is non availability or missing data that can be excluded according to C 4.5 classification algorithm. In this case additional information (x) can be added with training data to improve information gain and gain ratio of C4.5. We named this proposed algorithm as M C 4.5. In the case of two N/A values the probability of cause of disease rate is high to old age class and hence the value of C_1 must be updated with x value. Let x be the additional information added to A (training attribute for splitting), then the value is added to C_1. The updated class value based on the additional information is $C_1 = 5$ respectively.

$$SplitInfo_A(D) = \sum_{j=1}^{v \geq x} \frac{|D_j|}{|D|} \times Info(D_j)$$

$$\mathrm{x} = \mathrm{I}_A(D) = |\left(\mathrm{I}_{A1}(\mathrm{D}) + \mathrm{I}_{A2}(\mathrm{D})\right)|$$
$$\ldots\ldots\ldots + |\mathrm{I}_{An}(\mathrm{D}))| = |\mathrm{v}|$$
$$\{\text{only if } \mathrm{A} \geq \mathrm{I}(\mathrm{D})\}$$

Here the existing value of $I_A(D) = 4$ and the splitting criterion of A must be changed due to Additional information of x. Therefore

$$\mathrm{x} = \mathrm{I}_A(4) = |\left(\mathrm{I}_{A1}(4) + \mathrm{I}_{A2}(1)\right)|$$
$$\ldots\ldots\ldots + |\mathrm{I}_{An}(\mathrm{D}))| = |\mathrm{v}|$$

Therefore the revised information gain based on M C 4.5 algorithm is

$$I(D) = -\frac{5}{7}\log_2(\frac{5}{7}) - \frac{1}{7}\log_2(\frac{1}{7})$$

$$I(D) = 0.747bits$$

Therefore gain ratio of M C 4.5 is calculated as

$$G(male) = I(D) - \mathrm{I}_{male}(D)$$
$$= 0.747 - 0.523 = 0.224bits$$

6. Performance Evaluation

From the example. It is proved that M C 4.5 achieves more information gain and Gain ratio with mixed attributes. In this approach we repeat the same process with information gain and gain ration of C 4.5 with M C 4.5 respectively, but we do not freeze the test sets, we instead reduced the overall information gain by increasing the information gain of training attribute (I_A) and gain ratio. By this proposed methodology we also handled missing values and transformed as a valued attribute to improve the performance of C 4.5. The Table show

the Information gain, the percentage of M c 4.5 is 50% more when compared with C 4.5 decision tree classifier.

Table (2) Information Gain of C 4.5 and M 4.5

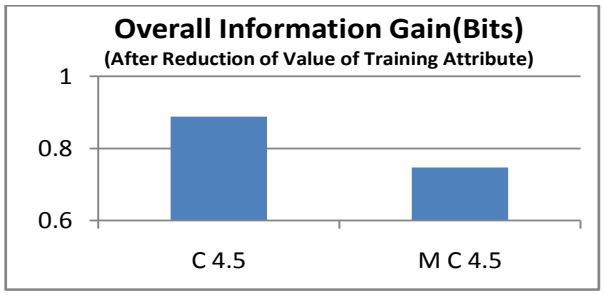
Decision Tree	Information Gain(Bits)
C 4.5	0.888
M C 4.5	0.747

Similarly, in the case of gain ratio of C 4.5 and M C4.5 attains more than more than 60 % double amount of gain ratio is achieved.

Table (3) Gain Ratio of C 4.5 and M 4.5

Decision Tree	Gain Ratio(Bits)
C 4.5	0.365
M C 4.5	0.224

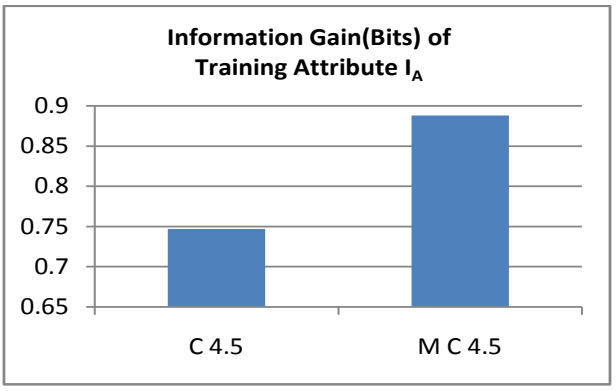
The following graphs show the difference of C 4.5 and M C.45 with respect to Information gain ratio. After selection of Training attribute A the MC4.5 obtains less score in overall information gain because the addition of special information gains more individually and it is reduced from overall information gain. The following figure shows the comparision between C 4.5 and M 4.5.



Fig(1). Comparision of Information Gain Value

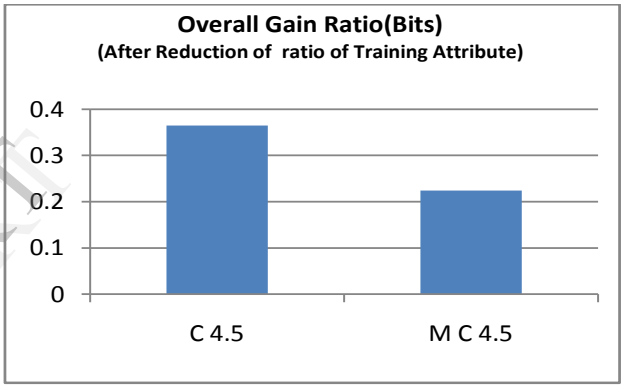
Note: The Lowest overall information gain shows high information is retained from training attribute and vice –versa.

The following figure shows the Information gain of Training Attribute of C 4.5 and M C4.5. The graph respresents highest information gain in M C4.5.



Fig(2).Information Gain of Training Attribute

The next figure shows the better results of overall gain ratio in M C4.5.It shows that C 4.5 gives high value (low potential information attained) because of less overall information gain.



Fig(3). Comparision of Gain Ratio

7. Conclusion and Future Work

In this paper, we presented a modified approach for C 4.5 algorithm based on its information theory and gain ratio. We also proposed the new methodology to handle missing valued attributes, adding of additional values to training sets .The example and implementation of proposed method gives better results for proposed M C4.5 algorithm. There are so many improvements developed by researchers in C 4.5 for various kinds of applications. We tried to develop the new modified C 4.5 algorithm for large distinct valued attributes. As a part of future work we propose another hybrid approach to handle multidimensional data with large intervals with the use of M C4.5 Algorithm

8. References

[1] Y. Yuan and M.J. Shaw, Induction of fuzzy decision trees. Fuzzy Sets and Systems 69 (1995), pp. 125–139.

[2]. IEEE Global History Network page about Nyquist criterion

[3].John R. Pierce and Rudy Kompfner K.J.Astrom: Nyquist and his seminal papers, 2005 presentation Nyquist biography, p. 2.

[4].Hartley, R.V.L., "The Function of Phase Difference in the Binaural Location of Pure Tones," *Physical Review*, Volume
Issue 6, pp 373–385, (June 1919).
[5].Hartley, R.V.L., Fry T.C.,"The Binaural Location of Pure Tones", *Physical Review*, Volume 18, Issue 6, pp 431 – 442, (December 1921).
[6].Hartley, R.V.L., "Relations of Carrier and Side-Bands in Radio Transmission", *Proceedings of the IRE*, Volume 11, Issue 1, pp 34 – 56, (February 1923).
[7].Hartley, R.V.L., "Transmission of Information", [1], *Bell System Technical Journal*, Volume 7, Number 3, pp. 535–563, (July 1928).
[8]. James, I. (2009). "Claude Elwood Shannon 30 April 1916 -- 24 February 2001". Biographical Memoirs of Fellows of the Royal Society 55: 257–265. doi:10.1098/rsbm.2009.0015.
[9]Bell Labs: Claude Shannon, the father of Information Theory, (1955).
[10].G. J. Chaitin, Algorithmic Information Theory, IBM Journal of Research and Development 21 pp. 350,359, 496. (1977).
[11].Quinlan, J. R., Discovering rules from large collections of examples: A case study, in D. Michie, ed., `Expert Systems in the Micro Electronic Age', Edin- burgh University Press. (1979)
[12].Quinlan, J. R., Induction of decision trees' Machine Learning 1, 81{106. Reprinted in Shavlik and Dietterich (eds.) Readings in Machine Learning, (1986).
[13].Quinlan, J. R., Inductive knowledge acquisition: A case study, in J. R. Quinlan, ed., `Applications of Expert Systems', Addison-Wesley, chapter 9, pp. 157, (1987).
[14].Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.
[15].Jaiwei Han and Micheline Kamber , Data Mining Concepts and Techniques.second Edition,Morgan Kaufmann Publishers. March 2006 publication.
[16] Chen Jin,Luo De –lin,mu Fen-xiang ,An Improved ID3 Decision tree algorithm,Xiamen University,2009.
[17].Rong Cao,Lizhen Xu,Improved C4.5 Decision tree algorithm for the analysis of sales.Southeast University Nanjing211189,china,2009.
[18].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.
[19].Surbhi Hardikar, Ankur Shrivastava and Vijay Choudhary Comparison betweenID3 and C4.5 in Contrast to IDS VSRDIJCSIT, Vol. 2 (7), 2012.
[20].Khalid Ibnal Asad, Tanvir Ahmed ,MD. Saiedur Rahman,Movie Popularity Classification based on Inherent, MovieAttributes using C4.5,PART and Correlation Coefficientd.IEEE/OSA/IAPR International Conference on Informatics, Electronics & Vision,2012
[21]. Gaurav L. Agrawal1, Prof. Hitesh Gupta2, Optimization of C4.5 Decision Tree Algorithm for Data Mining Application,IJTAE,ISSN 2250-2459, Volume 3, Issue 3, March 2013