Science and Technology Journal Vol. 7 Issue: 2 July 2019 ISSN: 2321-3388

# An Extended C4.5 Classification Algorithm using Mathematical Series

**R. Raja Aswathi[1]*, K. Pazhani Kumar[2] and B. Ramakrishnan[3]**
[1,2,3]Department of Computer Science, S.T. Hindu College, Nagercoil, India
E-mail: [1]*rajaaswathi@outlook.com, [2]skpk73@gmail.com, [3]ramsthc@gmail.com

**Abstract**—The algorithm C4.5 is an efficient decision tree based classification, which is derived from the ID3 approach. C4.5 is also a rule based classification algorithm. The main importance of the C4.5 algorithm is that it can deal with categorical data, over fitting of data and handling of missing values. The performance of C4.5 is superior to ID3 even with equal number of attributes. The EC4.5 (Exponential C4.5) is an extension of C4.5 algorithm which uses exponential of split value to predict the gain of attributes and handled the set back reported in C4.5. However the EC4.5 has some misclassification of data and to avoid this problem a new technique is introduced. This paper proposes a proficient technique TMC4.5 (Taylor-Madhava C4.5) to reduce the uncertainty in classification of data by integrating an exponential split value in EC4.5 and sin splitting value derived from the Madhava series. By using this technique an optimized gain value is obtained that reduces uncertainty. From the obtained result the TMC4.5 has far better results than the C4.5 and EC4.5 algorithms.

**Keywords:** *Data mining, Decision Tree, ID3, C4.5, TMC4.5.*

## INTRODUCTION

Data mining is the analysis of useful patterns from a large existing database. The data mining model is of two types as predictive and descriptive. The predictive model predicts the data using known results that are gathered from historical data. The tasks of the data mining predictive model include classification, regression, time series analysis, and prediction. The descriptive model identifies the classes using the similarity of patterns or relationship in data. Clustering, sequence discovery, summarization and association rules are descriptive in nature (Dunham, 2003).

This paper concentrates on the predictive model based classification methodology like the decision tree based algorithms. Decision tree, as the name implies, is a predictive model that can be viewed as a tree structure, where specifically each branch of the tree is a classification question and the leaves of the tree are partitions of the dataset with their classification (Clayson *et al.* 2006; Hussina *et al.* 2014). The two important steps in the decision tree technique that are most common in practice are to built the tree and then apply the tree to the existing database. Some of the decision tree algorithms are Classification and Regression Tree (CART), Chi-Squared Automatic Interaction Detection (CHAID), C4.5 or J48, Scalable Parallelizable Induction of Decision Trees (SPRINT). TheEC4.5 algorithm has many advantages over the ID3 algorithm as it is an extension of C4.5, despite of that it gives almost equivalent results when the attributes are same in number. However, in this paper the performance of the predictive EC4.5 is further enhanced by incorporating exponential and sin technique of the Taylor-Madhava Series with the gain ratio. As a result an exponential and sin modification of gain is proposed.

## RELATED WORK

ID3 algorithm was introduced by Quinlan Ross .It is based on Hunt's algorithm and the algorithm is serially implemented (Idriss *et al*. 2019). The ID3 algorithm is uses a general classification function to predict the splitting attributes, and it has many advantages, such as understandable decision rules and the intuitive model (Gaganjot *et al*. 2014). It also has some of its own disadvantages, for example: (1) the existence of a problem of multi-value bias on the process of attribute selection (Fayyad *et al*. 1992); (2) it is not easy to calculate information entropy (Liang *et al*. 2008; Exarchos *et al*. 2007) by using logarithmic algorithms, which increases the time of execution; and (3) the tree size is difficult to control (Quinlan, 1987), and the ID3 algorithm cannot handle large datasets with categorical attributes which

https://doi.org/10.22232/stj.2019.07.02.06

results in a bigger tree generation. The ID3 approach uses entropy values of the attributes to predict the information gain of the attributes.

## Entropy

ID3 calculates the entropy value using the given by the below equation,

$$Entropy\ (P) = \sum_{i=1}^{s}((pi)\log\left(\frac{1}{pi}\right))$$

## Information Gain

The information gain is calculated from the following formula,

$$Gain(P,A) = Entropy\ (P) - \sum_{i=1}^{s}((pi)\ Entropy\ (pi))$$

## METHODOLOGY

To overcome the uncertainty in EC4.5 the Taylor-Madhava series is calculated along with the split information value.

## DATA COLLECTION

The dataset contains the results of heart disease dataset which is used to compare the uncertainty level of the algorithms. This dataset uses 9 attributes with 303 instances which are represented in numerical formats.

**Table 1: Represents the Dataset Variables Format and Type (UCI Machine Learning Repository, 1988)**

| S. No | Attribute Name | Format |
|---|---|---|
| 1. | Chest Pain type (cp) | 1: Typical angina<br>2: Atypical angina<br>3: Non-anginal pain<br>4: Asymptomatic |
| 2. | Cholesterol(chol) | 168, 178, . . . . . |
| 3. | Fasting Blood Sugar > 120 mg/dl (fbs) | (1 = true; 0 = false) |
| 4. | Resting Electrocardiographic Results (restecg) | 0: normal<br>1: having ST-T wave abnormality<br>2: showing probable or definite left ventricular hypertrophy by Estes' criteria |
| 5. | Exercise Induced Angina (exang) | (1 = yes; 0 = no) |
| 6. | Slope of the peak exercise ST segment (slope) | 1: up sloping<br>2: flat<br>3: down sloping |
| 7. | Number of major vessels coloured by fluoroscopy (ca) | (0-3) |
| 8. | thal | 3 = normal; 6 = fixed defect;<br>7 = reversable defect |
| 9. | Diagnosis of Heart Disease (Target) | 0: < 50% diameter narrowing<br>1: > 50% diameter narrowing |

## EXISTING METHODOLOGIES

### C4.5

The C4.5 decision tree algorithm is an improved version of the ID3 algorithm. It uses the gain ratio value to predict the splitting attribute, where as the ID3 uses information gain value to determine the splitting attribute.

Consider the probability distribution (P = p1, p2, p3, . . . pi) and D denotes the dataset.

The entropy value of P is,

$$Entropy\ (P) = \sum_{i=1}^{s}(pi\log\left(\frac{1}{pi}\right)) \quad (1)$$

The information gain is calculated as,

$$Gain\ (P, A) = Entropy\ (P) - \sum_{i=1}^{s}((pi)Entropy\ (pi)) \quad (2)$$

The split information of the dataset D is,

$$SplitInfo\ (D) = \sum_{i=1}^{s}(|Dj|\ /|D|)\ \log(|Dj|/|D|) \quad (3)$$

The split information value is calculated for all attributes in the dataset D. The Gain (P, A) is divided by the equation (3) to get the GainRatio (D, A),

$$GainRatio\ (D, A) = \frac{Entropy\ (P) - \sum_{i=1}^{S}((pi)\ Entropy\ (pi))}{SplitInfo\ (D)} \quad (4)$$

### C4.5 Algorithm

**Input:** Database D

```
Tree = Null
If D is empty OR has no more attributes to split then
Terminate
End if
for all attributes where a∈D do
Calculate the GainRatio for 'a'
End for
$a_{high}$ = Attribute with highest gain value
```

```
9.  Tree= Create a decision node with $a_{high}$ in the root
10. $D_s$ = Reduced Sub-Dataset from D based on the $a_{high}$ attribute
11. for all $D_s$ do
12.     $Tree_s$ = C4.5 ($D_{s)}$
13.     Attach the $Tree_s$ to the Tree based on the attribute values
14. End for
15. Return Tree
```

## EC4.5 (Exponential C4.5)

In the exponential C4.5, the SplitInfo (D) is replaced by β as

$$GainRatio\ (D, A) = \frac{Entropy\ (P) - \sum_{i=1}^{s}((pi)\ Entropy\ (pi))}{\beta} \quad (5)$$

From the above equation (5), β is represented as,

$$\beta = \frac{Entropy\ (P) - \sum_{i=1}^{s}((pi)\ Entropy\ (pi))}{GainRatio\,(D,A)} \quad (6)$$

From the above equation if β = 1 then the gain value of ID3= C4.5. It is overcome by the Taylor series (Idriss *et al*. 2019). Consider the Taylor series for the exponential function $e^x$,

$$e^x = \sum_{n=0}^{\infty} x/n! \quad (7)$$

For x = 1 and n taking the limit, n ® ¥

$$\frac{\beta}{1!} + \frac{\beta}{2!} + \frac{\beta}{3!} + \ldots .\ \frac{\beta}{n!} \quad (8)$$

Which implies, n ® ¥ = $e^{\beta}$

$$E - split = e^{SplitInfo(D)} \quad (9)$$

That is equivalent to the format,

$$E\text{-}split = Exp\ (\ \sum_{i=1}^{s}(|Dj|\ /|D|)\ log(|Dj|/|D|)\ ) \quad (10)$$

By dividing the equation (2) by (10) the EC4.5 is analyzed (Idriss *et al*. 2019),

$$EC4.5 = \frac{Gain\ (P,A)}{E - split} \quad (11)$$

## EC4.5 Algorithm

**Input:** Database D

```
1.  Tree = Null
2.  If D is empty OR has no more attributes to split then
3.      Terminate
4.  End if
5.  for all attributes where $a \in D$ do
6.      Calculate the EC4.5 using exponential value for 'a'
7.  End for
8.  $a_{high}$ = Attribute with highest exponential split value
9.  Tree= Create a decision node with $a_{high}$ in the root
10. $D_s$ = Reduced Sub-Dataset from D based on the $a_{high}$ attribute
11. for all $D_s$ do
12.     $Tree_s$ = EC4.5 ($D_{s)}$
13.     Attach the $Tree_s$ to the Tree based on the attribute values
14. End for
15. Return Tree
```

## Proposed Methodology

### TMC4.5 (Taylor-Madhava C4.5)

In the Taylor-Madhava C4.5, the SplitInfo (D) is replaced by λ as,

$$GainRatio\ (D, A) = \frac{Entropy\ (P) - \sum_{i=1}^{s}((pi)\ Entropy\ (pi))}{\lambda} \quad (12)$$

As derived from the above equation λ is represented as,

$$\lambda = \frac{Entropy\ (P) - \sum_{i=1}^{s}((pi)\ Entropy\ (pi))}{GainRatio\ (D,A)} \quad (13)$$

(Idriss *et al*. 2019) From the above equation, if we consider the value of λ = 1. Then the value of Gain Ratio in C4.5 will be equal to the gain value in ID3. The limitation of ID3 is reflected in the Gain Ratio value. To overcome this problem the Taylor Madhava Series is used.

Consider the Taylor series for the exponential function $e^x$ at $a = 0$ is,

$$e^x = \sum_{n=0}^{\infty} x/n!$$

The Madhava Sin Series is represented as,

$$Sin\ x = x - \frac{x3}{3!} + \frac{x5}{5!} - \frac{x7}{7!} \quad (14)$$

For x = 1 and n taking the limit, n ® ¥

$$\frac{\lambda}{1!} + \frac{\lambda}{2!} + \frac{\lambda}{3!} + \ldots .\ \frac{\lambda}{n!} \quad (15)$$

And

$$1-\frac{\lambda}{1!}+\frac{\lambda3}{3!}-\frac{\lambda5}{5!}+\dots\dots\dots\dots+\frac{\lambda n}{n!} \quad (16)$$

If n ® ¥ the value of x®λ, which implies, $e^{\lambda}$ and sin λ. From equation (13), the value of λ is equivalent to SplitInfo(D). Now the split value is found by summing the exponential and sin split information as,

$$Split = e^{SplitInfo(D)} + Sin\ (SplitInfo(D)) \quad (17)$$

By dividing the gain value and information value the TMC4.5 is analyzed,

$$TMC4.5 = \frac{Gain\ (P,A)}{Split} \quad (18)$$

## TMC4.5 Algorithm

**Input:** Database D

```
Tree = Null
If D is empty OR has no more attributes to split then
    Terminate
End if
for all attributes where $a \in D$ do
    Run TMC4.5 (a)
    Set $a_{high}$ = Attribute with highest split value
    Calculate the split information using sin and exponential value for all 'a'
End for
Tree= Create a decision node with $a_{high}$ in the root
$D_s$ = Sorted Dataset D based on the $a_{high}$ attribute from TMC4.5 (a) function
Creating Tree from $D_s$ dataset
for all $a \in D_s$ do
    $Tree_s$ = TMC4.5 (a)
    Attach the $Tree_s$ to the Tree based on the attribute values
End for
Return Tree
```

## IMPLEMENTATION OF PROPOSED METHODOLOGY (TMC4.5)

Before using the entire dataset to predict the gain values, the first 10 records are examined with 9 attributed to find out the performance of algorithms. The algorithms that are subjected to this test are ID3, C4.5, EC4.5 and TMC4.5. The below datasets presented are used for the comparison of algorithm with best gain value.

The data provided in the about table are tested with the C4.5, EC4.5 and TMC4.5 algorithms to predict the gain values of the splitting attributes. And the outcome is represented in the below Table 3. The reduction in gain value of algorithm has the high rate of producing correct outcome and here the TMC4.5 has the reduced gain value and it can handle data without any uncertainty in classification.

**Table 2: Shows the Dataset with 10 Records and 9 Attributes**

| Dataset 1 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| ID | cp | Chol | Fbs | Restecg | Exang | Slope | ca | Thal | Target |
| 1 | 3 | 233 | 1 | 0 | 0 | 0 | 0 | 1 | 1 |
| 2 | 2 | 250 | 0 | 1 | 0 | 0 | 0 | 2 | 1 |
| 3 | 1 | 204 | 0 | 0 | 0 | 2 | 0 | 2 | 1 |
| 4 | 1 | 236 | 0 | 1 | 0 | 2 | 0 | 2 | 1 |
| 5 | 3 | 233 | 1 | 0 | 0 | 0 | 0 | 1 | 1 |
| 6 | 0 | 256 | 1 | 0 | 1 | 2 | 2 | 3 | 0 |
| 7 | 0 | 407 | 0 | 0 | 0 | 1 | 3 | 3 | 0 |
| 8 | 0 | 217 | 0 | 1 | 1 | 0 | 0 | 3 | 0 |
| 9 | 3 | 282 | 1 | 0 | 0 | 1 | 1 | 2 | 0 |
| 10 | 0 | 288 | 1 | 0 | 1 | 0 | 2 | 3 | 0 |

**Table 3: Represents the Uncertainty Level of C4.5, EC4.5 and TMC4.5 with 10 Records**

| Selected Attributes | C4.5 | EC4.5 | TMC4.5 |
|---|---|---|---|
| Exang | 0.4491 | 0.1639 | 0.1242 |
| Thal | 0.4438 | 0.14745 | 0.12105 |
| Cp | 0.3923 | 0.11432 | 0.0992 |
| Ca | 0.3882 | 0.1267 | 0.10496 |
| Chol | 0.3203 | 0.04407 | 0.04403 |
| Slope | 0.1609 | 0.05411 | 0.04415 |
| Restecg | 0.0395 | 0.01443 | 0.0109 |
| Fbs | 0.0290 | 0.0106 | 0.0081 |

The comparison of C4.5, EC4.5 and TMC4.5 are represented in the bellow Fig. 1

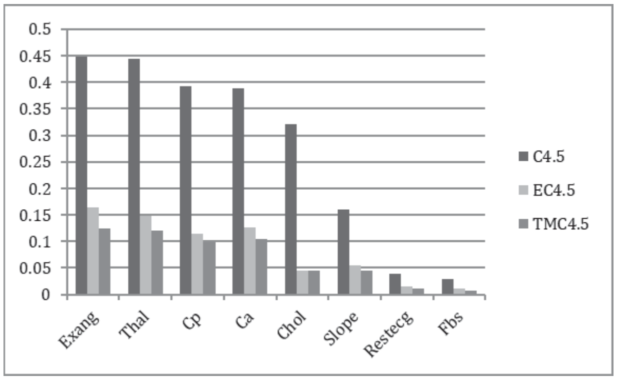

**Fig. 1: Represents the Uncertainty Level of the Algorithms**

## RESULTS AND DISCUSSION

From the implementation process of 10 records the TMC4.5 algorithm produces better results and now the 303 records are processed in the same way for efficient output. Here the Gain Ratio value of C4.5, EC4.5 and TMC4.5 are compared to find the probability of uncertainty in the given dataset. The below table represents the comparison of gain ratio values.

**Table 4: Represents the Uncertainty Level of C4.5, EC4.5 and TMC4.5**

| Selected Attributes | C4.5 | EC4.5 | TMC4.5 |
|---|---|---|---|
| Thal | 0.1646 | 0.0583 | 0.0464 |
| Exang | 0.1560 | 0.0571 | 0.0433 |
| Cp | 0.1176 | 0.0359 | 0.0306 |
| Ca | 0.1116 | 0.0351 | 0.0295 |
| Slope | 0.0902 | 0.0320 | 0.0253 |
| Chol | 0.0794 | 0.00486 | 0.00483 |
| Restecg | 0.0221 | 0.0081 | 0.00624 |
| Fbs | 0.0093 | 0.0030 | 0.0023 |

The following figure shows the uncertainty level of C4.5, EC4.5 and TMC4.5 algorithms. Here the TMC4.5 results in very less uncertainty as compared to the other two algorithms with 303 records.

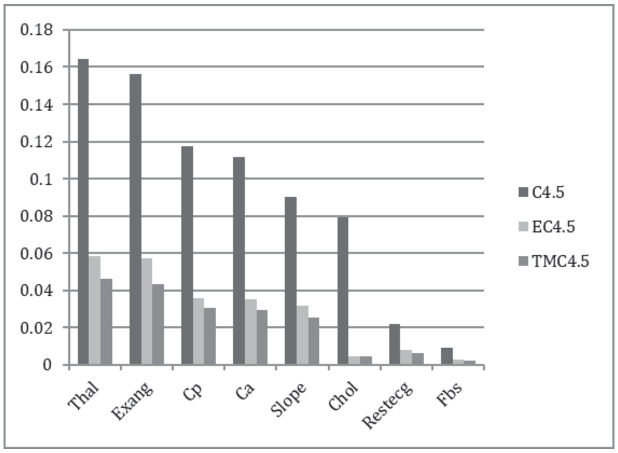

**Fig. 2: Represents the Uncertainty Level of the Algorithms and TMC4.5 has Very Less Level of Uncertainty**

## CONCLUSION

This paper proposes an improved version of the algorithm C4.5. The TMC4.5 uses the information theory of entropy and an enhanced gain ratio value to find the best splitting attribute using the sin and exponential split value. The best splitting attribute is identified as per the TMC4.5 algorithm to reduce the uncertainty in the classification of data. From the predicted results the TMC4.5 algorithm produces an optimized results and very low possibility of uncertainty in predicted data.

## REFERENCES

Clayson D, Sheffet MJ (2006) Personality and the Student Evaluation of Teaching, Journal of Marketing Education, vol. 2, no. 28, pp. 149-160.

Dunham MH (2003) Data mining: Introductory and advanced topics, Pearson Education, 4-5.

Exarchos TP, Tsipouras MG, Exarchos CP, Papaloukas C, Fotiadis DI, Michalis LK (2007) A methodology for the automated creation of fuzzy expert systems for ischaemic and arrhythmic beat classification based on a set of rules obtained by a decision tree. Artif. Intell. Med, 40, 187–200.

Fayyad UM, Irani KB (1992) The Attribute Selection Problem in Decision Tree Generation. In Proceedings of the National Conference on Artificial Intelligence, San Jose, CA, USA, 12–16 July 1992, pp. 104–110.

Gaganjot K, Amit C (2014) Improved J48 Classification Algorithm for the Prediction of Diabetes, International Journal of Computer Application, vol. 98, no. 5, pp. 13-17.

Hussina B, Merbouha A, Ezzikouri H (2014) A Comparative Study of Decision Tree ID3 and C4.5, International Journal of Advanced Computer Science, vol. 3, no. 1, pp. 13-19.

Idriss S, Lawan A (2019) An Improved C4.5 Model Classification Algorithm Based on Taylor's Series, JJCIT; (0):1.

Liang J, Shi Z (2008) The Information Entropy, Rough Entropy and Knowledge Granulation in Rough Set Theory, Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 2008, 12, 37–46.

Quinlan JR (1987) Generating Production Rules from Decision Trees. In Proceedings of the International Joint Conference on Artificial Intelligence, Cambridge, MA, USA, 23–28 August 1987; pp. 304–307.

UCI Machine Learning Repository (1988) Heart Disease Dataset, http://archive.ics.uci.edu/ml/datasets/Heart+Disease