

AN IMPROVED C4.5 MODEL CLASSIFICATION ALGORITHM BASED ON TAYLOR'S SERIES

I. I. Sinam¹ and Abdulwahab Lawan²

(Received: 3-Jan.-2019, Revised: 25-Feb.-2019, Accepted: 11-Mar.-2019)

ABSTRACT

C4.5 is one of the most popular algorithms for rule base classification. Many empirical features in the algorithm exist, such as continuous number categorization, missing value handling and over-fitting. However, despite its promising advantage over the Iterative Dichotomiser 3 (ID3), C4.5 has the major setback of presenting the equivalent result as the ID3, especially when the same number of attributes is used. This paper proposes a technique that will handle the setback reported in C4.5. The performance of the proposed technique is measured based on better accuracy. The Entropy of Information Theory is measured to identify the central attribute for the dataset. The researchers apply exponential splitting information (EC4.5) in utilizing the central attribute of the same dataset. The result obtained on introducing Taylor series suggested a far better result than when the C4.5 (gain ratio) was introduced.

KEYWORDS

ID3 Algorithm, C4.5 Algorithm, Information gain, Entropy, Gain ratio.

1. INTRODUCTION

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 [1]-[2]. It is a logical model represented as a binary or multiclass tree that shows how the value of a target variable can be predicted by using the values of a set of predictor variables. Decision tree classifiers are considered “white box” classification models, as they can provide the explanation for their classification models and can be used directly for decision -making [3]. Many decision tree algorithms exist and these include: Alternating Decision Tree (LAD), C4.5 or J48 Pruned Tree, Classification and Regression Tree (CART), Chi-squared Automatic Interaction Detection (CHAID), Quest, ...etc. Decision tree algorithms such as C4.5 had been developed earlier and continue to be regularly used in solving everyday classification tasks. However, despite its promising advantage over the ID3 algorithm, C4.5 has the major setback of presenting the equivalent result as the ID3, especially when the same number of attributes is used. In this paper, the predictive performance of this algorithm is enhanced by proposing another technique that will handle the noticeable setback and even present a more promising result than the C4.5 using (gain ratio). It is on this background that the exponential modification of the gain ratio is being proposed.

2. RELATED WORK

ID3 tree algorithm was introduced in 1986 by Quinlan Ross. It is based on Hunt's algorithm and the algorithm is serially implemented. The ID3 uses an information gain measure in choosing the splitting attribute [4]. The basic strategy in ID3 is the selection of splitting attributes with the highest information gain. That is; the amount of information associated with an attribute value that is related to the probability of occurrence [2]. Once the attribute has been chosen, the amount of information is measured, which is known as entropy [5]. Entropy is used to measure the amount of uncertainty, surprise or randomness in a dataset. The entropy will be zero when all the data in the set belong to the single class. One of the challenges with this approach is when ID3 selects the attribute having more number of values, which may necessarily not be the best attribute [5]. When testing a small sample, data may be over-fitted or over-classified. At a time, only one attribute is used for the testing purpose. As specified

1. I. I. Sinam is with Department of Computer Science, Bayero University, Kano, Nigeria.

2. A. Lawan is with Department of Information Technology, Bayero University, Kano, Nigeria. Email: abd_wahhb@yahoo.com

above, continuous data is difficult to analyze, as many trees need to be generated to find the perfect place to split the data, which makes the algorithm computationally expensive. The mathematical model of C4.5 is given by Equation (1).

$$Gain(P, A) = Entropy(P) - \sum_{j=1}^v ((p_j) Entropy(p_j)) \quad (1)$$

On the other hand, C4.5 algorithm is an extension of ID3 algorithm. It has an enhanced method of tree pruning that reduces miss-classification errors due to noise or too much detail in the training dataset found in ID3. It uses the gain ratio impurity method to evaluate the splitting attribute [2], [6]. Quinlan Ross introduced split information to information gain of ID3 as an improvement to overcome the limitations of ID3, which are latency and over-fitting and it becomes computationally expensive in handling continuous data. The gain ratio is given by Equation (2).

$$Gain_{Ratio}(D,A) = \frac{Entropy(D) \sum_{j=1}^i (p_j * Entropy(p_j))}{Splitting_{Info}} \quad (2)$$

- i. It will increase the performance when the number of attributes differs.
- ii. It will increase the performance when the number of attributes is the same.
- iii. And it will decrease the percentage of uncertainty in C4.5 algorithm.

3. METHODOLOGY

To overcome the limitations of C4.5, the researchers used Taylor's Series to modify the splitting information of C4.5.

3.1 Data Collection

The study uses an existing instructor's performance dataset from Abubakar Tafawa Balewa University Bauchi, Nigeria. The data collected was cleaned, normalized and organized in a form suitable for data mining process using WEKA platform. Table 1 shows the data format used for the research.

Table 1. Data format.

S/N	Variable Name	Variable Format	Variable Type
1.	Gender	Male, Female	Categorical
2.	Appointment Status (Appt. Status)	Permanent, Temporary, Contract	Categorical
3.	Employment Status (Emp. Status)	Old, New	Categorical
4.	Rank	Professor, Reader, SL,L1,AL, GA	Categorical
5.	Age	25, 30,...	Numerical
6.	University Working Experience (Univ. Exp.)	Year	Numerical
7.	Academic Qualification (Aca. Qual)	PhD, Master, Bachelor	Categorical
8.	Year of the Last Qualification	1996,1997	Numerical
9.	Professional Qualification (Prof. Qual.)	Yes, No	Categorical
10.	Average Unit Load	10, 15, 20, 24....	Numerical
11.	Formative Assessment Points	1, 2, 3....	Numerical
12.	Summative Assessment Point (SAP)	1, 2, 3....	Numerical
13.	Weighted Max Point (WmaxP)	20, 30, 40, 50.....	Numerical
14.	Performance	Satisfactory, Average, Poor	Categorical

The data consists of both categorical and numerical data making it suitable to perform this experiment.

3.2 The Existing Model (C4.5)

The C4.5 algorithm is an improvement of the ID3 algorithm, developed by Quinlan Ross in 1993. It uses gain ratio as an extension of gain information of ID3.

3.2.1 Mathematical Model (C4.5)

Let's consider the following probability distribution ($P = p_1, p_2, p_3, p_4, \dots, p_v$) and a dataset D and define the information carried by the distribution otherwise known as the entropy of P, proposed by [14]-[15], [18] given as:

$$Entropy(P) = -\sum_{j=1}^v p_j \log_2(p_j) \quad (3)$$

And the gain information for a test A is given by:

$$Gain(P, A) = Entropy(P) - \sum_{j=1}^v ((p_j) Entropy(p_j)) \quad (4)$$

We can define the splitting information in the form:

$$SplitInfo_A(D) = -\sum_{j=1}^v \frac{|D_j|}{|D|} \log_2 \left(\frac{|D_j|}{|D|} \right) \quad (5)$$

Let us consider a dataset D of some certain attributes with element $a_1, a_2, a_3, \dots, a_n$, where the gain ratio of such data set is given by:

$$Gain_{Ratio(D,A)} = \frac{Entropy(D) \sum_{j=1}^i (p_j * Entropy(p_j))}{SplittingInfo} \quad (6)$$

The two limitations associated with ID3; i.e., latency and over-fitting error are being improved by the gain ratio. The algorithm of C4.5 is shown below.

3.2.2 Algorithm of C4.5

Input: an attribute-valued dataset D

1. Tree = { }
2. If D is "pure" OR other stopping criteria met then
3. terminate
4. end if
5. for all attribute $a \in D$ do
6. Compute the gain ratio if we split on a
7. end for
8. a_{best} = Best attribute according to the above-computed criteria
9. Tree = Create a decision node that tests a_{best} in the root
10. D_v = Induced Sub-dataset from D based on a_{best}
11. for all D_v do
12. $Tree_v = C4.5(D_v)$
13. Attach $Tree_v$ to the corresponding branch of the Tree
14. end for
15. return Tree

3.3 The Proposed Model (EC4.5)

Suppose that we replace the split inform; i.e., denominator in Equation (6) with β

$$Gain_{Ratio(D,A)} = \frac{Entropy(D) \sum_{j=1}^i (p_j * Entropy(p_j))}{\beta} \quad (7)$$

The gain ratio is known to present a better result than the information gain if the set of element $a_i \neq a_j$, but if $a_i = a_j$, the result of the gain ratio and information gain is the same. We can see that from (4) if $\beta = 1$:

$$GainRatio(D, A) = Entropy(D) \sum_{j=1}^i (p_j Entropy(p_j)) \quad (8)$$

Equation (7) shows that when split info (β) =1, then ID3 = C4.5.

To overcome this:

If we let β be the subject, we can rewrite (7) as:

$$\beta = \frac{Entropy(D) \sum_{j=1}^i (p_j * Entropy(p_j))}{Gain_Ratio(D, A)} \quad (9)$$

Now, from (7), for $\beta =1$

Consider a Taylor's series

$$1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots \frac{x^n}{n!} \quad (10)$$

For $x = 1$, the series can be rewritten as:

$$\frac{\beta}{1!} + \frac{\beta}{2!} + \frac{\beta}{3!} + \dots \frac{\beta}{n!} \quad (11)$$

By taking the limit at $n \rightarrow \infty$,

$$\frac{\beta}{1!} + \frac{\beta}{2!} + \frac{\beta}{3!} + \dots \frac{\beta}{n!} \quad n \rightarrow \infty = e^\beta \quad (12)$$

e^β is called optimal split information; therefore, it optimizes the splitting information by splitting the value away from (1). It works for both cases: when $a_i = a_j$ and when $a_i \neq a_j$. The new technique suggests the introduction of a new parameter to the splitting information. We denote this $E - split$ and it is defined in the form:

$$E - split = e^{[SplitInfo_A(D)]} \quad (13)$$

This is equivalently defined as:

$$E - split = Exp \left(\sum_{j=1}^v \frac{|D_j|}{|D|} \log_2 \left(\frac{|D_j|}{|D|} \right) \right) \quad (14)$$

The introduction of the new parameter suggests that the splitting values are spread around the value 1. This helps in obtaining a better result. The division of Equation (2) by Equation (5) leads to the new method $E.C4.5$ which it is defined as:

$$E.C4.5 = \left(\frac{Gain(P, A)}{E - split} \right) \quad (15)$$

3.3.1 Algorithm of the Proposed EC4.5

Input: an attribute-valued dataset D

1. Tree = { }
2. If D is "pure" OR other stopping criteria met then
3. terminate
4. end if
5. for all attribute $a \in D$ do
6. Compute the gain ratio using exponential split if we split on **a**
7. end for
8. a_{best} = Best attribute according to the above computed criteria
9. Tree = Create a decision node that tests a_{best} in the root
10. D_v = Induced Sub-dataset from D based on a_{best}
11. for all D_v do
12. $Tree_v = EC4.5(D_v)$

13. Attach $Tree_v$ to the corresponding branch of the Tree
14. end for
15. return Tree

3.4 Evaluation

We consider the following terms in evaluating the performance of the proposed EC4.5.

- (a) TN (True Negative) is the number of correct predictions that an instance is invalid.
- (b) FP (False Positive) is the number of incorrect predictions that an instance is valid.
- (c) FN (False Negative) is the number of incorrect prediction that an instance is invalid.
- (d) TP (True Positive) is the number of correct predictions that an instance is valid.

Also, the following performance measure was used to test the performance of the proposed EC4.5. Accuracy is the proportion of the total number of predictions that were correct:

$$Accuracy(\%) = \frac{(TN+TP)}{(TN+FN+FP+TP)} \quad (16)$$

Precision is the proportion of the predicted valid instances that were correct:

$$Precision(\%) = \frac{TP}{(FP + TP)} \quad (17)$$

Recall is the proportion of the valid instances that were correctly identified:

$$Recall(\%) = \frac{TP}{(FN + TP)} \quad (18)$$

F-Measure is derived from precision and recall values:

$$F - Measure(\%) = \frac{(2 \times Recall \times precision)}{(Recall + Precision)} \quad (19)$$

The F-Measure is used because despite the Precision and Recall values are valid metrics in their own right, one of them can be optimized at the expense of the other. The F-Measure only produces a high result when Precision and Recall are both balanced and significant.

4. IMPLEMENTATION OF THE PROPOSED MODEL (EC4.5)

To see how the new model works, we consider the following example. Suppose that we want to compare the performances of ID3, C4.5 and EC4.5 to decide whether the time will be good to play basketball. A two-week data collection was used.

Table 2. Experimental dataset [19].

Day	Outlook	Temperature	Humidity	Play
1	Sun	Hot	High	No
2	Sun	Hot	High	No
3	Overcast	Hot	High	Yes
4	Rain	Sweet	High	Yes
5	Rain	Cold	Normal	Yes
6	Rain	Cold	Normal	No
7	Overcast	Cold	Normal	Yes
8	Sun	Sweet	High	No
9	Sun	Cold	Normal	Yes
10	Rain	Sweet	Normal	Yes
11	Sun	Sweet	Normal	Yes
12	Overcast	Sweet	High	Yes
13	Overcast	Hot	Normal	Yes
14	Rain	Sweet	High	No

The classification of the target is "Should we play basketball?" The answer can be either yes or no. The weather attributes which include outlook, temperature and humidity take the following values:

Outlook = {Sun, Overcast, Rain}

Temperature = {Hot, Sweet, Cold}

Humidity = {High, Normal}

So, using the three now sets: the information gain (ID3), the gain ratio (C4.5) and the E-gain ratio (EC4.5) are calculated for the outlook based on temperature and humidity as shown in the appendix.

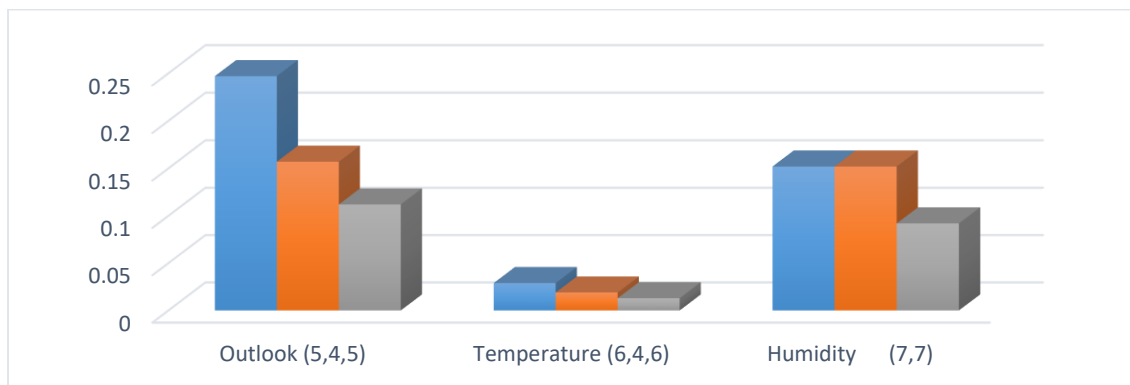


Figure 1. Outcome of the 3 classification algorithms.

From Figure 1, the three classification algorithms ID3, C4.5 and EC4 have the following outcome: outlook with 5,4,5 attributes shows that ID3 has a value of 0.247, C4.5 has a value of 0.157 and EC4.5 has a value of 0.112. Subsequently, temperature with 6,4,6 attributes shows that ID3 has a value of 0.029, C4.5 has a value of 0.019 and EC4.5 has a value of 0.013. However, humidity which has the same number of attributes of 7,7 leads to ID3 and C4.5 having the same value of 0.152. EC4.5 shows an improvement by having the value 0.092 which reduces the number of uncertainties in C4.5.

5. RESULTS AND DISCUSSION

In the experiment, the values of the gain ratio (C4.5) and E-gain ratio (EC4.5) were first used to calculate the probability of uncertainty of some selected attributes with the highest instances. The outcome is shown in detail in Table 3. Figure 2 displays that EC4.5 is the optimal algorithm which has the lowest probability of uncertainty on all attributes and C4.5 has the highest probability of uncertainty.

Table 3. Probability of uncertainty outcome of gain ratio and E-gain ratio.

Selected Attribute	C4.5 (Gain Ratio)	EC4.5 (E-Gain Ratio)
SAP	0.931464	0.335235
Aca. Qual.	0.639899	0.227053
Rank	0.429688	0.114575
Univ. Exp.	0.148147	0.082564
Age	0.090222	0.030221
Prof. Qual.	0.310135	0.110069
Appt. Status	0.105773	0.02568
Gender	0.002881	0.001252
Emp. Status	0.000418	0.000313

Furthermore, C4.5 and EC4.5 classification algorithm, were trained and tested on the same dataset; the measures used for the algorithm performance evaluation were accuracy, precision, recall and F1 measure.

Table 4 illustrates the detailed results of the two classification algorithms.

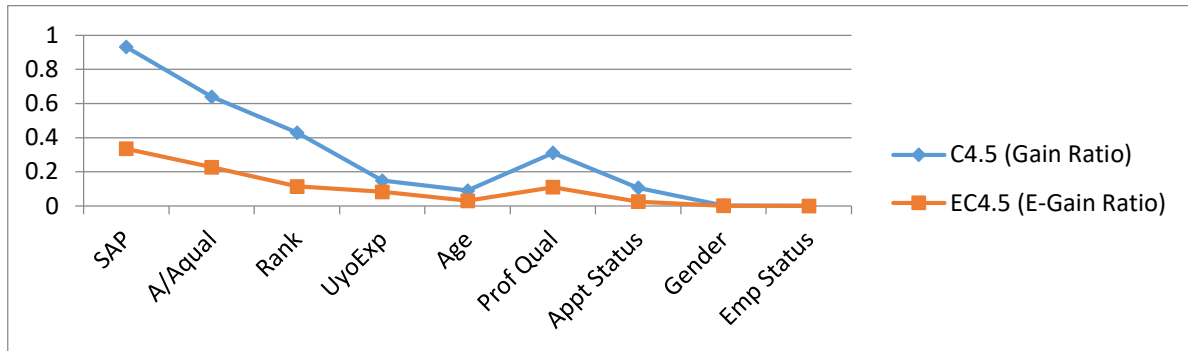


Figure 2. Probability of uncertainty outcome of gain ratio and E-gain ratio.

Table 4. Detailed classification accuracy results of C4.5 and EC4.5.

Performance Metrics	C4.5	EC4.5
Accuracy	51.27%	99.40%
Error Rate	48.73%	0.60%
Precision	0.513	0.994
Recall	0.513	0.994
F1 Measure	0.678	0.994

The detailed classification accuracies suggest that EC4.5 outperformed C4.5, because it has a lower FP rate of 0.003 and a TP rate of 0.994 which was used to calculate the accuracy using the performance metrics. Thus, EC4.5 is the optimal model of classification algorithm in this paper.

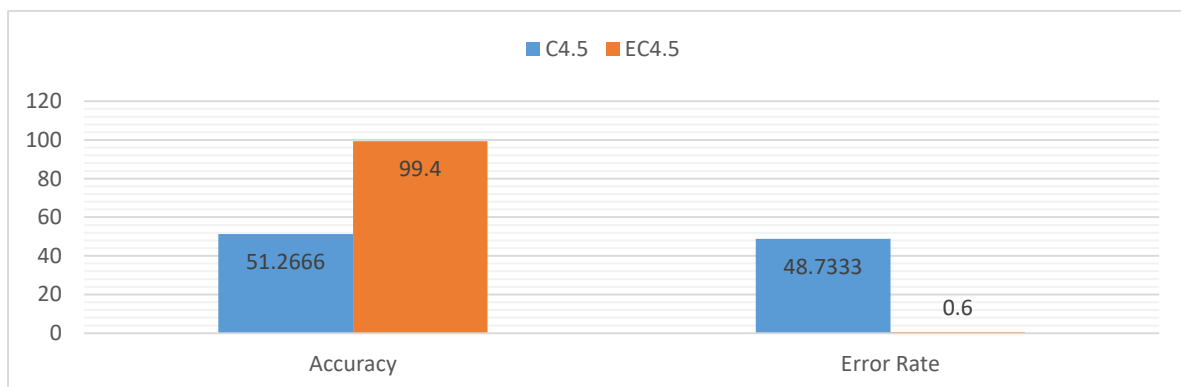


Figure 3. Accuracy and error rate of C4.5 and EC4.5.

Figure 3 shows that EC4.5 has the highest accuracy of 99.40% with an error rate of 0.60%, while C4.5 has an accuracy of 51.27% with an error rate of 48.73%.

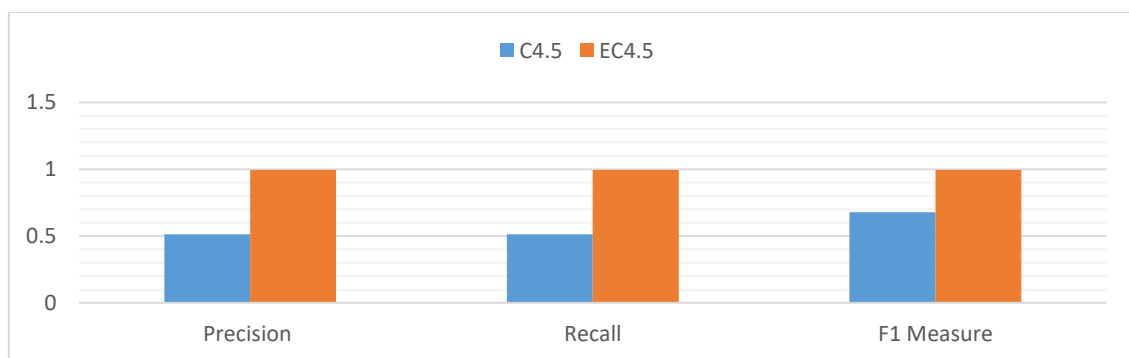


Figure 4. Precision, recall and F1 measure of C4.5 and EC4.5.

Figure 4 shows the detailed results of the compared algorithms, with C4.5 having the highest value over

EC4.5; under precision C4.5 has 0.513 and EC4.5 has 0.994; under recall C4.5 has 0.513 and EC4.5 has 0.994; and lastly under F1 measure C4.5 has 0.678 and EC4.5 has 0.994. The overall result suggested that EC4.5 is the optimal algorithm compared to C4.5.

6. CONCLUSIONS

This paper proposed a modified model (EC4.5). The proposed modification offers solutions to the limitations associated with C4.5 in terms of presenting an equivalent result with ID3 when the same number of attributes is used. After testing the two classifiers (C4.5 and EC4.5), the result of the experiment shows that EC4.5 outperformed, with an accuracy of 99.40%, whereas C4.5 has an accuracy of 51.27%. Based on the result of this research, EC4.5 was selected as the optimal algorithm. Future work should consider a hybrid approach to handle multi-dimensional data with large intervals using EC4.5 algorithm.

ACKNOWLEDGEMENTS

The authors would like to thank the management of Bayero University, Kano, Nigeria for their moral encouragement and financial support.

REFERENCES

- [1] D. Clayson and M. J. Sheffet, "Personality and the Student Evaluation of Teaching," *Journal of Marketing Education*, vol. 2, no. 28, pp. 149-160, 2006.
- [2] B. Hussina, A. Merbouha and H. Ezzikouri, "A Comparative Study of Decision Tree ID3 and C4.5," *International Journal of Advanced Computer Science*, vol. 3, no. 1, pp. 13-19, 2014.
- [3] C. Romero, L. O. Juan and V. Sebastian, "A Meta-learning Approach for Recommending a Subset of White-box Classification Algorithms for Moodle Datasets," *Journal of Theoretical and Applied Information Technology*, vol. 6, no. 5, pp. 268-271, 2013.
- [4] K. Gaganjot and C. Amit, "Improved J48 Classification Algorithm for the Prediction of Diabetes," *International Journal of Computer Application*, vol. 98, no. 5, pp. 13-17, 2014.
- [5] M. M. Mazid and A. Shawkat, "Improved C4.5 Algorithm for Rule-based Classification," *Proceedings of the 9th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases*, pp. 296-301, 2010.
- [6] A. Moustafa, E. Shahira and M. Essam, "Defining Difficult Laryngoscopy Findings by Using Multiple Parameters, A Machine Learning Approach," *Egyptian Journal of Anaesthesia*, vol. 33, no. 2, pp. 153-158, 2017.
- [7] Y. Yuan and M. Shaw, "Fuzzy Sets and Systems," *Elsevier*, vol. 69, no. 2, pp. 125-139, 1995.
- [8] A. B. Adeyemo and O. Oriola, "Personal Audit Using a Forensic Mining Technique," *International Journal of Computer Science*, vol. 7, no.7, pp. 222-231, 2010.
- [9] K. Asha, A. M. Gowda and M. Jayaram, "Comparative Study of Attribute Selection Using Gain Ratio and Correlation-based Feature Selection," *International Journal of Information Technology and Knowledge Management*, vol. 2, no. 2, pp. 271-277, 2010.
- [10] G. Chaitn, "Algorithmic Information Theory," *Journal of Research and Development*, vol. 8, no. 4, pp. 350-359, 2000.
- [11] S. Hardikar, A. Shrivastava and V. Choudary, "Comparison between ID3 and C4.5," *International Journal of Computer Science*, vol. 2, no. 7, pp. 34-39, 2012.
- [12] L. Gaurav and G. Hiaesh, "Optimization of C4.5 Decision Tree Algorithm for Data Mining Application," *International Journal of Information Technology and Knowledge Management*, vol. 3, no. 3, pp. 2250-2459, 2013.
- [13] R. Hartley, "The Function of Phase Difference in the Binaural Location of Pure Tones," *Journal of Advanced and Applied Sciences*, vol. 13, no. 6, pp. 373-385, 2000.
- [14] S. Kumar and E. Ramaraj, "Modified C4.5 Algorithm with Improved Information Entropy and Gain Ratio," *International Journal of Engineering Research and Technology*, vol. 2, no. 9, pp. 2768-2773, 2013.
- [15] K. Santhosh, "Modified C4.5 Algorithm with Improved Information Entropy," *International Journal of*

- Engineering Research & Technology, vol. 2, no. 14, pp. 485-512, 2013.
- [16] W. Yisen, S. Chaobing and X. Shu-Tao, "Improving Decision Trees by Tsallis Entropy Information Metric Method," Proc. International Joint Conference on Neural Networks, Vancouver, BC, Canada, IEEE Xplore, 24 – 27, July, 2018.
- [17] M. M. Mazid, "Improved C4.5 Algorithm for Rule-based Classification," in Mastorakis Nikos (ed.), Artificial Intelligence Knowledge Engineering and Database, vol. 7, no. 5, pp. 296-301, 2017.
- [18] A. Neeraj, G. Bhargava and M. Manish, "Decision Tree Analysis on J48 Algorithm for Data Mining," International Journal of Advanced Research in Computer Science, vol. 3, no. 6, pp. 22-45, 2013.
- [19] M. Dragon and G. Lujbisa, "The Use of Data Mining for Basketball Matches Outcome Prediction," Proc. of the 8th IEEE International Symposium on Intelligent Systems and Informatics, pp. 309-312, Serbia, 2010.
- [20] I. H. Witten, E. Frank, L. Trigg, M. Hall, G. Holmes and S. J. Cunningham, "Weka: Practical Machine Learning Tools and Techniques with Java Implementations," (Working Paper 99/11), Department of Computer Science, University of Waikato, Hamilton, New Zealand, vol. 31, pp. 76-81, 2000.
- [21] I. Al-Turaiki, M. Alshahrani and T. Almutairi, "Building Predictive Models for MERS-COV Infections Using Data Mining Techniques," Journal of Infection and Public Health, vol. 9, no. 6, pp. 744-748, 2016.
- [22] L. Yi-bin, W. Ying-ying and R. Xue-wen, "Improvement of ID3 Algorithm Based on Simplified Information Entropy and Coordination Degree," Proc. of Chinese Automation Conference, IEEE Xplore, Jinan, China, vol. 1, no. 3, pp. 88-92, 2017.
- [23] G. Attilio and N. Filippo, "Search-Intensive Concept Induction," International Journal of Computer Science, vol. 7, no. 6, pp. 137-145, 2000.
- [24] T. G. Kumar, "Advanced Applications of Neural Networks and Artificial Intelligence," International Journal of Information Technology, vol. 2, no. 6, pp. 57-68, 2012.
- [25] S. Mardikyan and B. Badur, "Analyzing Teaching Performance of Instructors Using Data Mining," Journal of Informatics in Education, vol. 10, no. 2, pp. 245-257, 2011.

APPENDIX

Result of the three classification algorithms

Outlook			Temperature			Humidity		
Algorithms	Attributes	Calculated Values	Algorithms	Attributes	Calculated Values	Algorithms	Attributes	Calculated Values
ID3	5,4,5	0.247	ID3	6,4,6	0.029	ID3	7,7	0.152
C4.5		0.157	C4.5		0.019	C4.5		0.152
EC4.5		0.112	EC4.5		0.013	EC4.5		0.092

ملخص البحث:

تعد الخوارزمية (C4.5) من أكثر خوارزميات التصنيف شيوعاً. وعلى الرغم من أفضليتها على الخوارزمية (ID3)، فهي تعطي نتيجة مماثلة لتلك التي يتم الحصول عليها باستخدام الخوارزمية (ID3)، وبخاصة عند استخدام العدد نفسه من السمات. وتقتصر هذه الورقة تقنية تعالج هذا القصور في الخوارزمية (C4.5). وقد تم تقييم أداء التقنية المقترحة بناءً على الدقة.

يستخدم الباحثان معلومات الفصل الأسي (EC4.5) للاستفادة من السمة المركزية لمجموعة البيانات نفسها، ويقدمان نموذجاً محسناً لخوارزمية التصنيف (C4.5) بناءً على سلسلة تايلور. وقد أعطى النموذج المحسن المقترح نتيجة أفضل بكثير مقارنة بنتائج الخوارزمية التقليدية (C4.5) تدل عليها الدقة الأعلى للنموذج المقترح والبالغة 99.4% مقارنة بدقة الخوارزمية الأصلية (C4.5) البالغة 51.27%.

