Journal of Engineering and Applied Sciences 14 (1): 253-258, 2019
ISSN: 1816-949X

Experimental Process of Data Laboratory Recruitment Using C4.5 Algorithm

Yasella Dina Aprilia, Roswan Latuconsina and Tito Waluyo Purboyo
Department of Computer Engineering, Faculty of Electrical Engineering, Telkom University, Bandung, Indonesia

Abstract: Data processing use C4.5 algorithm with Excel is useful to speed up manual calculations. Use of MS. Excel to help users to learn in understanding data mining algorithms before using programming languages like Java, C ++, R programming, etc. Data train is the data recruitment assistant laboratory. This data has several stages: file collection interview, written test, coding test internship test, teaching test, a coding test for 12 h. Based on the phase stage will be done with the calculation C4.5 algorithm. Based on the results of this experiment will form a decision tree.

Key words: C4.5 algorithm, phase stage recruitment, speed up, calculation, results, experiment

INTRODUCTION

Data mining is the process of finding a new pattern from very large data sets. In large data such as in banks insurance, operator companies, employee recruitment, it will be difficult to read data conventionally. Therefore, data mining is useful to speed up work (Aprilia *et al*., 2018). The extraction process or unusual data will be difficult to understand because it requires data mining to process the data. Data mining is very useful for decision cases that are very important in life (Connolly and Begg, 2010). Information and very large data usage will have special patterns and rules (Han and Kamber, 2006) as a large database the data mining process becomes a large relational database (Krishnaiah *et al*., 2013).

For some descriptive features used by the class is the classification method by generalizing the structure that will be applied to the new data (Hssina *et al*., 2014). To change the world of data mining based on MIT technology review (Larose, 2005). A very important method for calculating assessing and classifying data .

Based on the problems that exist in data processing are manual and slow. Processing with Excel makes it easy to use formulas and speed up work. The advantages of Excel are (Shi, 2011):

- In Excel for single factor analysis of variance steps
- Automation solution

This study conducts experiments on how data mining is done with MS. Excel. Data recruitment is derived from data recruitment laboratory assistant in the study program computer system. This data has several stages: file collection interview, written test, a coding test internship test, teaching test, coding test for 12 h.

MATERIALS AND METHODS

C4.5 algorithm: In data mining research now, there are several methods such as Bayesian, decision tree and so on. The most efficient classification method is the decision tree (Li *et al*., 2009). Examples of decision trees are C4.5 and ID3 algorithms. A decision tree is a method that is fast in predicting a case. Based on the lack of ID3, there is the development of the C4.5 algorithm (Khedr *et al*., 2016) which chooses the biggest gain of each attribute (Yao and Xing, 2011). An example of the application of the C4.5 algorithm is predicting student achievement. The results of the case reached an accuracy of 80-84% to determine bad behavior and provide special guidances for these students (Li *et al*., 2015).

To start calculating the C4.5 algorithm that is by calculating entropy. Entropy is a parameter to measure diversity in a set of data. The greater the diversity of data, the greater the entropy score. The formula to find entropy:

$$\text{Entropy}(S) \equiv \sum_{i=1}^{c} -p_i . \log 2 p_i \quad (1)$$

Where:

Corresponding Author: Yasella Dina Aprilia, Department of Computer Engineering, Faculty of Electrical Engineering, Telkom University, Bandung, Indonesia

c = The number of scores contained in target attribute
pi = The portion or ratio between the number of samples in class i and the number of samples in the data set

Information gain is the effectiveness of an attribute in classifying data. The formula to find information gain:

$$\text{Gain}(S, A) \equiv \text{Entropy}(S) \sum_{v \in \text{values}(A)} \frac{|S_v|}{|S|} \text{Entropy}(S_v) \quad (2)$$

Where:

A = Attribute
V = A possible score for attribute A
Scores (A) = The set of possible scores for attribute A
$|S_v|$ = Number of Samples for score v
$|S|$ = Total data Sample
Entropy (S_v) = Entropy for Samples that have a score of v

Information gain will experience problems for the attributes that have very varied scores, to solve this problem can be calculated by split information which is formulated as follows (Mitchell, 1997):

$$\text{Split Info}(S, A) \equiv \sum_{i=1}^{c} -\frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|} \quad (3)$$

Where:

S = A set of data samples
S_i until S_c = Subdivided sample data based on the number of variations in the attribute A

The gain ratio is formulated as follows (Mitchell, 1997):

$$\text{Gain ratio}(S, A) \equiv \frac{\text{Gain}(S, A)}{\text{Split information}(S, A)}$$

Decision trees make shapes like trees (Bao and Guan, 2016). The upper part is called the root node in the middle called the inner root and the bottom is called leaf. This data has several stages: file collection interview, written test, a coding test internship test, teaching test, coding test for 12 h (Fig. 1).

	A	B	C	D	E	F	G	H
26								
27	Files	interview	written test	coding test	internship test	teaching test	coding test for 12 hours	result
28	collect	v	71	60	72.7875	52	57.5	yes
29	collect	no	0	0	0	0	0	no
30	collect	v	57.2	100	86.1875	48	76.25	yes
31	collect	v	59	20	62.125	74	56.25	yes
32	collect	v	50	30	76.3125	50	43.75	yes
33	collect	v	64	30	41.325	38	71.875	yes
34	collect	v	66	20	75.7375	42	62.125	yes
35	collect	v	55	30	67.0625	40	50	no
36	collect	v	0	0	0	0	0	no
37	collect	v	0	0	0	0	0	no
38	collect	v	56	40	72.725	66	50	yes
39	collect	v	42	20	69.5375	39	40	yes
40	collect	v	73	40	80.075	56	82.125	yes
41	collect	v	81	65	83.8875	48	52.5	yes
42	collect	v	0	0	0	0	0	no
43	collect	v	74	60	82.4375	60	41.25	yes
44	collect	v	29.2	30	4.375	62	60.5	no
45	collect	v	59	30	57.625	40	50	no
46	collect	v	64	30	66.075	74	60.5	yes
47	collect	v	0	0	0	0	0	no
48	collect	v	44	0	0	0	0	no

Fig. 1: Recruitment phase

File collection stage: At this stage, the prospective assistant collects files such as CV, essays, photos, recaps, etc. If the prospective assistant completes the file, it will be written collecting. If there is one that is not listed then incomplete.

Interview test: Is conducted to determine the reasoning and leadership of the assistant candidate. Prospective assistants will be given a few questions. Question concerns cases that occur with the assistant. From the case will arise the problem and must be resolved. The way of speaking, appearance is also an aspect of assessment at this stage. There are 3 interviewers and 1 assistant candidate in each room so, the assessment is not subjective. It's just that in this data is not included the score of only the absenteeism of the prospective assistant.

Written test: At this writing test stage, the candidate is given some questions about the algorithm. There are two types of questions: multiple choice and essay. Each question number has a score. The score in the measurement in the C4.5 algorithm is divided into two namely the score of the assistant candidate who ≤50 and the score of the candidate's assistant >50.

Coding test: At this stage, the assistant candidate is to do the coding. This stage is the stage of how long assistants can know how far the prospective assistant understands the algorithm. There are several cases that the assistant candidate should solve using the algorithm. The score in the measurement in the C4.5 algorithm is divided into two namely the score of the assistant candidate who ≤50 and the score of the candidate's assistant >50.

Internship test: At this stage, the assistant candidate conducts an internship in the relevant laboratory. This internship stage is useful for prospective assistants more familiar with the laboratory. The internship lasts approximately one month. Assistant candidates will be assessed from liveliness, attendance and how to interact with other assistants. The score in the measurement in the C4.5 algorithm is divided into two namely the score of the assistant candidate who ≤50 and the score of the candidate's assistant >50.

Teaching test: Teaching this test to find out how a potential assistant faces a practitioner. Various simulations will be performed by several assistants from other laboratories. Teaching the test also to find out how deep the material is mastered by a prospective assistant. Prospective assistants must know all the rules of practice. The score in the measurement in the C4.5 algorithm is divided into two namely the score of the assistant candidate who ≤50 and the score of the candidate's assistant >50.

Coding test for 12 h: This stage of the assistant candidate performs a coding test for 12 h. Assistant candidates are divided into groups. From the group will be shown how the assistant candidate can work with other assistant candidates. The score in the measurement in the C4.5 algorithm is divided into two namely the score of the assistant candidate who ≤50 and the score of the candidate's assistant >50.

The process of doing the C4.5 algorithm:

- Select an attribute as root
- Create a branch for each score
- Divide the case in the branch
- Repeating the process of each branch until all the cases on the branch have the same class

Make a name in the cell area:

- Cell A28 until cell A48 is named files
- Cell B28 until cell B48 is named interview
- Cell C28 until cell A48 is named written
- Cell D28 until cell A48 is named coding
- Cell E28 until cell A48 is named internship
- Cell F28 until cell A48 is named teaching
- Cell G28 until cell A48 is named 12 h
- Cell H28 until cell A48 is named result

The determination of this branch candidate is useful for what attributes are used. The first attribute is the file. There are two assessments of collecting and not collecting. The second attribute is the interview there are two assessments that attend the interview or not. The third stage until the 7th assessment used the same score is ≤50 and >50 (Table 1-3).

The calculation of node 1 is the entropy of the total case. Frequency: number of the resultant of the yes case and the resultant number of a case no:

- Sum : the sum of cases yes and no
- P_j : the frequency divided by the sum
- $Log_2 P_j$: log2 of P_j
- $-P_j.log_2P_j$: -Pj multiplied by $log2P_j$

Table 1: Recruitment phase

Phases	A	B	C	D	E	F	G	H
27	Files	Interview	Written	Coding	Intern	Teaching	12 h	Result
28	√	√	71	60	72	52	57	yes
29	√	-	0	0	0	0	0	no
30	√	√	57.2	100	86	48	76	yes
31	√	√	59	20	62	74	56	yes
32	√	√	50	30	76	50	43	yes
33	√	√	64	30	41	38	71	yes
34	√	√	66	20	75	42	62	yes
35	√	√	55	30	67	40	50	no
36	√	√	0	0	0	0	0	no
37	√	√	0	0	0	0	0	no
38	√	√	56	40	72	66	50	yes
39	√	√	42	20	69	39	40	yes
40	√	√	73	40	80	56	82	yes
41	√	√	81	65	83	48	52	yes
42	√	√	0	0	0	0	0	no
43	√	√	74	60	82	60	41	yes
44	√	√	29.2	30	4	62	60	no
45	√	√	59	30	57	40	50	no
46	√	√	64	30	66	74	60	yes
47	√	√	0	0	0	0	0	no
48	√	√	44	0	0	0	0	no

Table 2: Determination of branch candidates

----------Determination----------	
Collection files	Not collection
Attend interview	Not attend interview
The score of written test ≤50	The score of written test >50
The score of coding test ≤50	The score of coding test >50
The score of internship test ≤50	The score of internship test >50
The score of teaching test ≤50	The score of teaching test >50
The score of coding test for 12 h≤50	The score of coding test for 12 h≤50

Table 3: Calculation of node 1 (root)

Phases	A	C	D	E	F
40	Results	Frequency	P_j	$Log2P_j$	$-P_jlog2P_j$
41	Yes	12	0.571428571	0.807354922	0.46134567
42	No	9	0.428571429	1.222392421	0.523882466
43	Sum	21			0.985228136

This result is the end result of entropy.

Algorithm 1; Frequency in Excel:

Formula frequency in Excel = COUNTIFS(result, B41)
Formula P_j in Excel = C41/C43
Formula Log_2 P_j in Excel =Log(D41, 2)
Formula Log_2 $-P_j.log_2P_j$ in Excel = -E41*D41

The formula calculates the number of yes and no cases in each attribute are = COUNTIFS (files, D46, result, E45). This table generates the entropy of each attribute:

- P (yes): number of yes cases divided by all cases of each attribute
- P (not): number of cases not shared by all cases of each attribute

Table 4: Calculate the number of yes and no cases in each attribute

Phases	B	C	D	E	F
45	Branch			yes	no
46	1	File collection	Collecting	12	9
47		File collection	Not collecting	0	0
48	2	Interview test	Present	12	8
49		Interview test	Not present	0	1
50	3	Written test	≤50	2	7
51		Written test	>50	10	2
52	4	Coding test	≤50	8	9
53		Coding test	>50	4	0
54	5	Internship test	≤50	1	7
55		Internship test	>50	11	2
56	6	Teaching test	≤50	6	8
57		Teaching test	>50	6	1
58	7	Coding test for 12 h	≤50	4	8
59		Coding test for 12 h	>50	8	1

- -P (yes) log 2 P (yes) :-P (yes) multiplied by log 2 P (yes)
- -P (not) log 2 P (not) : -P (not) multiplied by log 2 P (no)
- Entropy : The sum of the branches of each attribute

Algorithm 2; Formula P (yes):

The formula P(yes) in Excel based on Table 4 = IFERROR((E46/(E46+F46)), 0)
The formula P(not) in Excel based on Table 4 = IFERROR((F46/(E46+F46)), 0)
The formula P(-P (yes) log2 P (yes)) in Excel based on Table 4 = IFERROR((-G46*LOG(G46, 2)), 0)
The formula -P (not) log2 P (not) in Excel based on Table 4 =IFERROR((-H 46*LOG(H46, 2)), 0)
The formula entropy in Excel based on Table 4 =I46+J46

After calculating the entropy of each attribute, hence can be calculated gain.

Algorithm 3; Formula $|S_v|/|S|$ in Excel:

The formula $|S_v|/|S|$ in Excel based on Table 5 and 6 = COUNTIFS(files,D46)/COUNTA(files)
The formula Entropy * $|S_v|/|S|$ in Excel based on Table 5 and 6 = K46*L46
The formula sum entropy * $|S_v|/|S|$ each attribute in Excel based on Table 5 and 6 = SUM(M46:M47)
The formula Gain each attribute in Excel based on Table 5 and 6 = F43-N46

The highest gain will be node 1. After the gain calculation, it can be seen that the internship test has the highest gain that is 0.394728894. After knowing the highest gain, the decision tree can be made with the internship test is at the top. The test internship has two branches namely the score ≤50 and >50. In Fig. 2 can be explained if the score ≤50 then the number of yes = 1, the number of no = 11. If the score is >50, then the number of yes = 7, the number of no = 2.

Table 5: The total entropy of each attribute

Attributes		G	H	I	J	K
45	Branch	P(yes)	P(no)	-P(yes) log2 P(yes)	-P(no) log2 P(no)	Entropy
46	Collecting	0.57	0.42	0.46	0.52	0.98
47	Not collecting	0	0	0	0	0
48	Present	0.6	0.4	0.44	0.52	0.97
49	Not present	0	1	0	0	0
50	Written test ≤50	0.22	0.77	0.48	0.28	0.76
51	Written test >50	0.83	0.16	0.21	0.43	0.65
52	Coding test ≤50	0.47	0.52	0.51	0.48	0.99
53	Coding test >50	1	0	0	0	0
54	Internship test ≤50	0.12	0.87	0.37	0.16	0.54
55	Internship test >50	0.84	0.15	0.20	0.41	0.61
56	Teaching test ≤50	0.42	0.57	0.52	0.46	0.98
57	Teaching test >50	0.85	0.14	0.19	0.40	0.59
58	Coding test for 12 h ≤50	0.333	0.66	0.52	0.38	0.91
59	Coding test for 12 h >50	0.88	0.11	0.15	0.35	0.50

Table 6: Calculating the gain

Attributes		L	M	N	O
45	Branch	$\|S_v\|/\|S\|$	Entropy* $\|S_v\|/\|S\|$	Sum entropy* $\|S_v\|/\|S\|$ each attribute	Gain
46	Collecting	1	0.98	0.98	0
47	Not collecting	0	0		
48	Present	0.95	0.92	0.92	0.06
49	Not present	0.04	0		
50	Written test <=50	0.42	0.32	0.69	0.28
51	Written test >50	0.57	0.37		
52	Coding test ≤50	0.80	0.80	0.80	0.17
53	Coding test >50	0.19	0		
54	Internship test ≤50	0.38	0.20	0.59	0.39
55	Internship test >50	0.61	0.38		
56	Teaching test ≤50	0.66	0.65	0.85	0.13
57	Teaching test >50	0.33	0.19		
58	Coding test for 12 h ≤50	0.57	0.52	0.74	0.24
59	Coding test for 12 h >50	0.42	0.21		

From the number of yes and no in the internship attribute then cannot be decided whether if the total score ≤50 then yes or no similarity with a score >50. Repeating the C4.5 algorithm stage is as follows:

- Select an attribute as root
- Create a branch for each score
- Divide the case in the branch
- Repeating the process of each branch until all the cases on the branch have the same class

RESULTS AND DISCUSSION

The result of this data processing is a decision tree. Can be seen in Fig 3. If test score of internship ≤50 then will be seen interview test, if not attend the interview then result no. If attend the interview it will be seen the score of the written test, if ≤50 then the result no but if the score is >50 then the result yes.

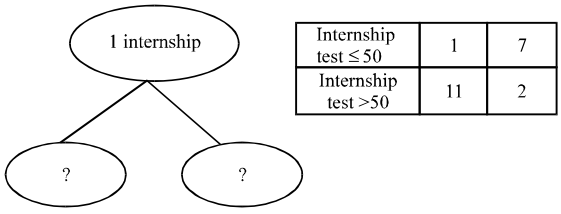

Fig. 2: Decision tree

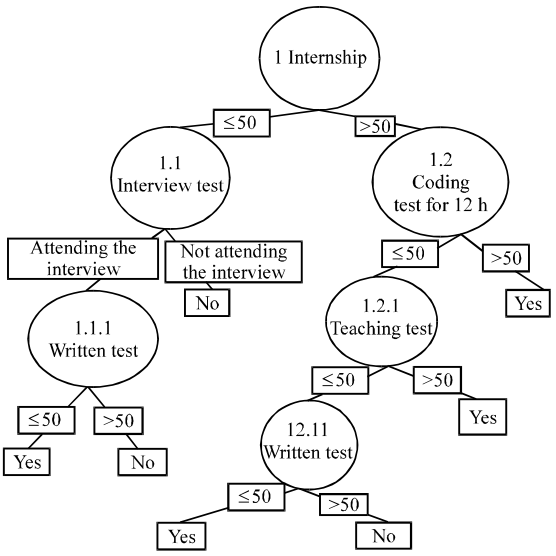

Fig. 3: The result of a decision tree

If the internship score >50 then it will be seen the coding test for 12 h if the score >50 then the result yes. If the score of internship test ≤50 will be seen the coding test for 12 h, if the score ≤50 then seen the score teaching test. If the score of teaching test >50 the result is yes but if the score ≤50 seen the score of written test. If the score written test ≤50 the result is yes but if the score >50 the result is no.

CONCLUSION

It can be seen that the C4.5 algorithm performed with Excel can speed up the time to quickly find out the recommendation of the assistant candidate in the next recruitment. So, it only takes some attributes.

With the development of programming languages such as C ++, Java, R programming it is better to use the programming language. For MS. Excel, it is better to

be used for beginners in learning algorithms because MS. Excel helps user's convenience and is known to everyone.

REFERENCES

Aprilia, Y.D., R. Latuconsina and T.W. Purboyo, 2018. A review of several algorithms for data mining. J. Eng. Appl. Sci., 13: 6157-6161.

Bao, X. and X. Guan, 2016. A method of predicting crude oil output based on RS-C4.5 algorithm. Proceedings of the 2016 3rd International Conference on Information Science and Control Engineering (ICISCE), July 8-10, 2016, IEEE, Beijing, China, ISBN:978-1-5090-2536-7, pp: 63-66.

Connolly, T.M. and C.E. Begg, 2010. Database Systems: A Practical Approach to Design, Implementation and Management. 5th Edn., Addison-Wesley, Boston, Massachusetts, USA., ISBN:9780321523068, Pages: 1243.

Han, J. and M. Kamber, 2006. Data Mining: Concepts and Techniques. 2nd Edn., Elsevier, Amsterdam, Netherlands, USA., ISBN:9780123739056, Pages: 770.

Hssina, B., A. Merbou, H. Ezzikouri, M. Erritali, 2014. A comparative study of decision tree ID3 and C4.5. Int. J. Adv. Comput. Sci. Applic., 2104: 13-19.

Khedr, A.E., A.M. Idrees and A.I. El Seddawy, 2016. Enhancing iterative dichotomiser 3 algorithm for classification decision tree. WIREs Rev. Data Mining Knowl. Dis., 6: 70-79.

Krishnaiah, V., D.G. Narsimha and D.N.S. Chandra, 2013. Diagnosis of lung cancer prediction system using data mining classification techniques. Int. J. Comput. Sci. Inf. Technol., 4: 39-45.

Larose, D.T., 2005. Discovering Knowledge in Data: An Introduction to Data Mining. John Wiley & Sons, Hoboken, New Jersey, USA., Pages: 222.

Li, L., S. Yao, Z. Ou and Q. Chen, 2015. Forecast of student achievement variation trend based on C4.5 decision tree. Proceedings of the International Conference on Artificial Intelligence and Industrial Engineering (AIIE 2015) Vol. 133, July 26-27, 2015, Atlantis Press, Paris, pp: 383-386.

Li, R., X.M. Wei and X.W. Yu, 2009. The improvement of C4.5 algorithm and case study. Proceedings of the 2nd International Symposium on Computational Intelligence and Design (ISCID'09) Vol. 2, December 12-14, 2009, IEEE, Changsha, China, ISBN:978-0-7695-3865-5, pp: 190-192.

Mitchell, T., 1997. Machine Learning. McGraw-Hill, New York.

Shi, H.B., 2011. Applying excel VBA to implement comparison among physical experiment data. Proceedings of the 2011 International Conference on Management and Service Science (MASS), August 12-14, 2011, IEEE, Wuhan, China, ISBN:978-1-4244-6579-8, pp: 1-3.

Yao, Y.F. and L.T. Xing, 2011. Improvement of C4.5 decision tree continuous attributes segmentation threshold algorithm and its application. J. Cent. South Univ. Sci. Technol., 42: 3772-3776.