IEESE International Journal of Science and Technology (IJSTE), Vol 2, Number 2 June, 2013,1-8
ISSN : 2252-5297

Prediction of Students Graduation Using Decision Tree Method with the Implementation of Algorithm C4.5

Susi Mashlahah[1], M. Ainul Yaqin[2], Muhammad Faisal[3]

[1.2.3] Informatics Engineering, Faculty of Science and Technology
Islamic State University (UIN) of Maulana Malik Ibrahim of Malang
Jl. Gajayana No. 50 Malang
Indonesia

E-mail: suu_sy@yahoo.com, a_yaqinov@yahoo.com, muhfais@yahoo.com

Abstract. Islamic State University of Maulana Malik Ibrahim of Malang is a public college in Indonesia that receives more increased quota of students year to year, but all of students are not able to graduate just in time in accordance with a study period pursued so it results in the accumulation of the number of students who are not graduated in accordance with the graduation periods. Based on that background, the research is conducted to create a system using the technique of classification. The data processing is used to predict unknown-yet class, that is the prediction of students' graduation. Technique of classification used is a decision tree with the implementation of algorithm C4.5.
Input used is an attribute of students'data, includingg the origin of the region, types of school, way to university entrance, the experience of pesantren, accumulative grade point average, and grade point average of each semester, from first semester up to fifth semester. That students' data are training sample data used in arranging the decision tree. Based on the testing that use graduated students' data from 2005 up to 2008, accuracy of compatibility in this system reaches 82,79%, so it can be used to predict unknown-yet students' graduation.

Key-Words: *Classification, Decision Tree, Algorithm C4.5*

1. Introduction

Entering a new school year, student quota accepted at a state university is more increased, but it is not all students are able to graduate just in time in accordance with a study period pursued so it results in the number of students becomes growing a lot. The number of graduate students and new ones who enter in every year is not balanced, thus it needs a system which is able to be used to predict student's graduation. This prediction system of student graduation needs available information to find out whether a student can graduate just in time or not. If students' graduation can be found out early, the academics can imply a policy to minimize the number of students who is hot graduated just in time in accordance with his/her study periods.

This system is created by implementing one of classification techniques in mining data that is the decision tree. Many algorithms developed to make decision tree, such as ID3, CART and C4.5 [7]. Decision tree has bigger complexity, because in algorithm C4.5, every value in an attribute is traced and processed to get entropy of each value that will be used to search purity size of each attribute stated with information gain. This investigation process will form a pattern in the form of decision tree[1].

2. Decision Tree

The decision tree is one of the strongest and known classification methods. Decision tree method changes the big fact becomes a decision tree which represents rule, the rule can easily be interpreted by humans. Decision tree also functions to explore data, find a hidden relationship between a number of input variable and a target variable[2].

The decision tree model consists of a set of rule to divide a number of heterogeneous population becomes smaller one (homogenous) by paying attention to the objective target. Objective target is usually classified certainly and decision tree model more focuses on the calculation of probability from each record towards the category or to classify records by grouping

them in one class. A decision tree can be built by implementing one of decision tree algorithms to model unclassified class data collections[4].

The concept of the decision tree is to change data becomes decision tree and decision rules.

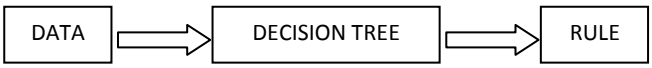

Figure 1. The Concept of Decision Tree

Data in the decision tree is usually stated in the form of a table with the attribute and record. Attribute states a parameter which is created as a criterion in the formation of tree. For instance, to determine playing tennis, criteria noticed are the weather, the wind, and the temperature. One of attributes is an attribute which states solution data per item called attribute target.

A main benefit from the use of a decision tree is its capability to break down complex decision taking process becomes simpler so the decision maker will more interpret solution of a problem. Decision Tree also functions to explore data, find a hidden relationship between a number of input variable candidate and a target variable. Decision tree fuses data exploration and modeling, so it is very good as an early step in modeling process even when it is created as the final model from some other techniques.

3. Algorithm C4.5

C4.5 is an algorithm used to produce decision tree developed by Ross Quinlan [1]. C4.5 is an extension of previous algorithm ID3. Decision tree produced by C4.5 can be used for classification, and for this reason, C4.5 is often called as classifier statistic[3]. Generally, algorithm C4.5 used to create decision tree are [4]:

a. Choose attribute as root.
b. Make branches for each value.
c. Divide class into branches.
d. Repeat process for each branch till all cases in branches have similar classes.

To choose attribute as root, it is based on the highest Gain value from available attributes. Meanwhile, to get Gain value, we must first calculate Entropy value of all values in the attributes. Entropy plays a role as a parameter to measure variance of a data sample. After entropy value in sample data is known, attribute effectiveness will be measured in classifying data, this size is called Gain information [9].

To accumulate gain, we use the formula as written in the following equation:

$$Gain\ (S, A) = Entropy\ (S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy\ (Si)$$

Notes:
S: a set of cases
A: attribute
N: the number of attribute A partition
|Si| : the number of cases in the-i partition
|S| : the number of cases in S

Entropy is used to determine how informative an attribute input to produce attribute output. The basic formula of the entropy is as follows:

$$Entropy\ (S) = \sum_{i=1}^{n} - p_i * log_2\, p_i$$

Notes:
S: a set of cases
n : the number of S partition
p_i : proportion of S_i towards S

4. Research Method

4.1 System Overview

Input from system that will be created is students data in the form of table containing several attributes such as students university entrance way data, the origin of school, the origin of region, accumulative grade point average, grade point average of the first semester up to fifth semester, the information of pesantren, entering year and graduation year of students. Criterion of on time graduation is the length of 4 years study counted from graduation year minus entering year of students. If it is more than 4 years, it includes in the classification of improper time graduation cases.

The students data will be classified based in target to be achieved that is whether students graduation can be on time or not, then the data is executed based on the procedure from decision tree method that is algorithm C4.5 to search entropy value and gain information. After calculation process is finished, it will result in the rule or condition used in the determination of decision in prediction process. Output of this system is the notification of on time graduation or not of each student predicted.

System created has following capabilities:

1. Calculating data obtained by changing its format into the form of training sample data table.
2. Calculating data to determine gain and entropy values.
3. Updating data by re-training new data sample.
4. Advising or giving new knowledge in the determination of target to be predicted. In this case, it refers to the students graduation time with study period pursued whether it is exact or not.

4.2 Instruments and Materials

a. Hardware Needs

A PC/laptop to perform design and creation of system with following specifications:

- Processor Core 2 Duo with Windows 7 Operational System
- Memory 2 GB

b. Software Needs

Besides hardware needs, the authors also need software needs to design and create system. The software needs are:

- Web Browser Google Chrome 24.0.1312.52 Version
- Power Designer 6.1 to design system
- Appserv 2.5.9 Version for web server
- My SQL 5.0.45 Version to save data
- Adobe Photoshop CS3 to design application appearance
- Microsoft Office 2010 to make documentation and report of research result

4.3 Research Place

The research is conducted in Islamic State University of Maulana Malik Ibrahim of Malang by using students data of Informatics Engineering major study, Faculty of Sciences and Technology that are graduated from 2005 to 2008.

4.4 System Flow

In figure 2, system flow to be created will be explained with diagram block.

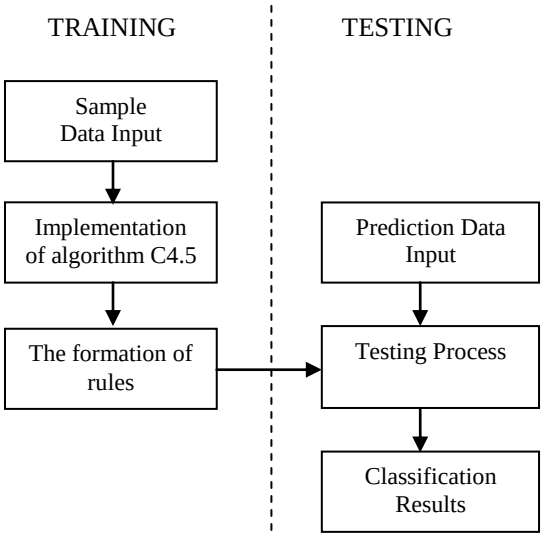

Figure 2. System Diagram Block

Diagram block can be explained that the system starts by two steps:

1. Training Process
   In the training process, it is performed by entering sample data into table prepared for calculation process. The table includes attributes, the number of overall data, the number of data classified based on the target determined, in this case, graduation on time or not, and the column of entropy and gain values. The next stage, the implementation of algorithm C4.5 formula, that is by the searching of Entropy and Gain values in each attribute to be used as form of tree. Tree is classification rule form which will be implemented in the testing process.

2. Testing Process
   In this testing process, the step performed is by entering testing data or prediction data. Attribute used in this testing process must be suitable with the attribute in the training process. Every attribute data will be compared with the rules which are already formed in the previous training data calculation. Then, attribute condition from testing data will be classified based on the target to be known that is the students prediction of whether they can graduate on time or not.

5. Results And Discussion

The trials were conducted three times using the number of different training sample data. The first one is by using 60 students data, the second one uses 79 data and the third one uses 90 data. Meanwhile, for testing data uses 93 students data of Informatics Engineering major study of Islamic State University of Maulana Malik Ibrahim of Malang. Testing data collection is taken from several students data of 2005 to 2008 who are already graduated. The trial results will be matched with the real data results whether the students graduate on time or not.

Prediction process in the application is conducted through these following stages:

1. Data is entered to the field of origin of the region, the origin of the school, way of university entrance, accumulative grade point average, grade point average of the first to fifth semesters, and pesantren. Attribute use is quite similar with training data input, the difference of both of them is the graduation notification. In the training data, graduation notification has been known in order to make the case classification can be counted by using *a Decision Tree method* to get prediction rules. Meanwhile, in testing data, graduation notification attribute has not been known and the graduation results will be predicted.

2. After being successful to be entered, each record of testing data attributes will be matched with the rules formed when training data calculation process.
3. If testing data entered has a similar record with the rules above, it will be classified into case 'Y' that is graduate on time. Meanwhile, if the record from attribute data entered is not similar, it will be classified into decision 'N', that is not graduating on time.

The table 1 produced in the calculation process using 79 students data.

Table 1. The Calculation of Entropy and Gain

NODE	ATTRIBUTE	ATTRIBUTE VALUE	S	Y	N	ENTROPY	GAIN
0	**TOTAL**		**79**	**9**	**70**	**0.511639784**	
	THE ORIGIN OF REGION						0.042559524
		WEST JAVA	1	0	1	0	
		EAST JAVA	66	7	59	0.487917993	
		BORNEO	1	0	1	0	
		MADURA	5	2	3	0.970950594	
		SUMATRA	6	0	6	0	
	THE TYPE OF SCHOOL						0.063865605
		MAN	20	1	19	0.286396957	
		MAS	14	1	13	0.371232327	
		SMAN	31	5	26	0.637387499	
		SMAS	10	1	9	0.468995594	
		SMKN	1	0	1	0	
		SMKS	3	1	2	0.918295834	
	THE WAY OF UNIVERSITY ENTRANCE						**0.121386942**
		ACHIEVEMENT INDEPENDENT	5	0	5	0	
		WRITTEN TEST INDEPENDENT	**49**	**2**	**47**	**0.246022578**	
		WRITTEN TEST SNMPTN	**16**	**3**	**13**	**0.69621226**	
		INVITATION SNMPTN	**8**	**3**	**5**	**0.954434003**	
		SPMB PTAIN	1	1	0	0	
	ACCUMULATIVE GPA						0.106076167
		A	2	2	0	0	
		B	65	7	58	0.49291578	
		C	12	0	12	0	
	GPA OF SEMESTER 1						0.080415815
		A	5	2	3	0.970950594	
		B	50	7	43	0.584238812	
		C	22	0	22	0	
		D	2	0	2	0	
	GPA OF SEMESTER 2						0.057624521
		B	57	9	48	0.629249224	
		C	22	0	22	0	
	GPA OF SEMESTER 3						0.089437905
		A	1	1	0	0	
		B	57	8	49	0.58515699	
		C	21	0	21	0	
	GPA OF SEMESTER 4						0.045528553
		A	3	0	3	0	
		B	61	9	52	0.60365225	
		C	15	0	15	0	
	GPA OF SEMESTER 5						0.028952697
		A	7	1	6	0.591672779	
		B	60	8	52	0.566509507	
		C	12	0	12	0	
	PESANTREN						0.030741192
		NO	50	8	42	0.634309555	
		YES	29	1	28	0.216396932	

Data is grouped based on the attribute and its attribute value then it is counted the total number, the number of students who graduate on time (Y) and not graduate on time (N), then it is counted entropy and gain values of each attribute.

The row of the TOTAL of ENTROPY column in the table above is counted by using the following formula:

$$Entropy\ (Total) = \left(-\frac{9}{79} * \log_2\left(\frac{9}{79}\right)\right) + \left(-\frac{70}{79} * \log_2\left(\frac{70}{79}\right)\right) Entropy(Total)$$
$$= 0.511639784$$

Meanwhile, GAIN value in the row of THE ORIGIN OF THE REGION is counted by using Gain formula, as following:

$$Gain\ (Total, The\ Origin\ of\ The\ Region) = Entropy(Total) - \sum_{i=1}^{n} \frac{|The\ Origin\ of\ The\ Region|}{|Total|} * Entropy(The\ Origin\ of\ The\ Region)$$

$$Gain\ (Total, The\ Origin\ of\ The\ Region) = 0.511 - \left(\left(\frac{1}{79} * 0\right) + \left(\frac{66}{79} * 0.487\right) + \left(\frac{1}{79} * 0\right) + 579*0.970+679*0\right)$$

$$Gain(Total, The\ Origin\ of\ The\ Region) = 0.0425$$

From the results of table above, it can be known that attribute with the highest Gain is THE WAY OF UNIVERSITY ENTRANCE, that is 0.1213. Therefore, THE WAY OF UNIVERSITY ENTRANCE becomes root node. There are 5 attribute values from THE WAY OF UNIVERSITY ENTRANCE, they are ACHIEVEMENT INDEPENDENT, WRITTEN TEST INDEPENDENT, WRITTEN TEST SNMPTN, INVITATION SNMPTN and SPMB-PTAIN. This is tree formed:

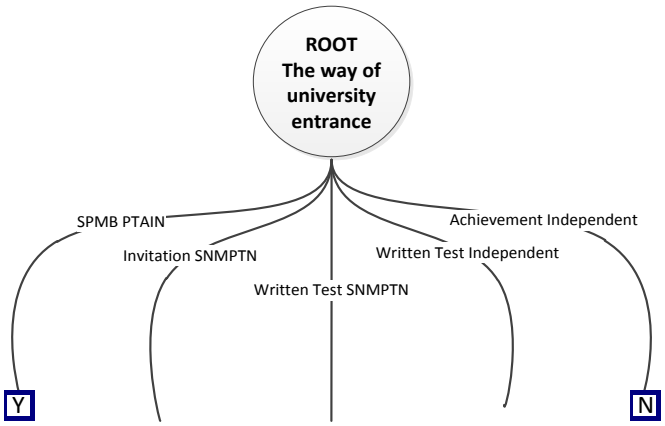

Figure 3. Root The way of University Entrance

Those five attribute values are classified based on Y and N values from attribute with the highest gain as following:

Table 2. Entropy gain The way of University Entrance

ATTRIBUTE	ATTRIBUTE VALUE	S	Y	N	ENTROPY	GAIN
THE WAY OF UNIVERSITY ENTRANCE						0.121386942
	ACHIEVEMENT INDEPENDENT	5	0	5	0	
	WRITTEN TEST INDEPENDENT	49	2	47	0.246022578	
	WRITTEN TEST SNMPTN	16	3	13	0.69621226	
	INVITATION SNMPTN	8	3	5	0.954434003	
	SPMB PTAIN	1	1	0	0	

If Y value = 0 and N has value, the attribute results in 1 rule that is Not Graduate On Time, while if N = 0 and Y has value, it will result in 1 Rule that is Graduate On Time. Based on the table above, attribute value of ACHIEVEMENT INDEPENDENT is known Y value = 0 and N

= 5 so it results in 1 Rule that is IF THE WAY OF UNIVERSITY ENTRANCE OF ACHIEVEMENT INDEPENDENT, THEN IT IS NOT GRADUATE ON TIME. Attribute value of SPMB PTAIN has Y value = 1 and N = 0 so it results in 1 Rule again that is IF THE WAY OF UNIVERSITY ENTRANCE OF SPMB PTAIN, THEN IS GRADUATE ON TIME. Meanwhile, for WRITTEN TEST INDEPENDENT, WRITTEN TEST SNMPTN, AND INVITATION SNMPTN have Y and N values so they need to be counted again to find the next Node.

In the rule process in the application, there are some repetitions in the formation of decision tree based on the acquisition of training data or accustomed data. In the first repetition, there are some functions, they are function which results in the highest gain and get Node function used to classify cases into rule of graduate on time case (Y), not graduate on time (N), or new Node. The second repetition has some functions, they are take-further and get Node-further. The difference is attribute with the highest gain which is resulted from the previous calculation, not entered into next calculation array.

Meanwhile, in the prediction process, attribute in the target data entered will be corrected with available training sample data calculation results rule. Whether the condition of attribute from the students' data entered has decided of graduate on time or not.

From that test, it can be known the degree of the truth by comparing target result of the applications with the real data. It can be seen as following:

Table 3 The test result of the degree of the truth

No	The number of training data	The number of rules	The number of testing data	The number of correct predictions	The number of wrong predictions	Truth accuracy (%)
1	60	17	93	60	33	64.51
2	79	19	93	66	27	70.96
3	90	22	93	77	16	82.79

The determination of accuracy level can be counted by using following formula:

$$\frac{\text{The Number Of Correct Prediction Data}}{\text{The Number Of Testing Data}} \text{ x } 100\%$$

From the testing result above, it shows that more training sample data used, more truth accuracy level of prediction. The number of prediction which is suitable with the real data is bigger than the number of unfit predictions, so the use of algorithm C4.5 as decision tree former in the prediction system of students' graduation can be used because it has proven its accuracy. The rules resulted from decision tree show the determination of case classification based on the target to be achieved that is decision of graduate on time or not.

6. Conclusion and Discussion

6.1 Conclusion

The conclusions obtained from the result of final assignment working about prediction system of students graduation using algorithm C4.5 are:

1. From testing result using 60 sample data, the pattern formed has match accuracy by 64.51%, while from the 79 sample data result in 70.96%, and from 90 sample data result in 82.79%.
2. More sample data used, the prediction will be more rightful, so algorithm C4.5 can be used to predict unknown class, that is by predicting students graduation can be on time or not.

6.2 Suggestions

Some suggestions from the authors to the further research development are:

1. Making input dynamically to add or change attributes and values inside which are used in the calculation process of algorithm C4.5.
2. Making Tree form visually on the attribute which has highest Gain value till the calculation process is finished in order to know the kinds of attributes positing root, branch and leaf.

Reference

[1] Suhartinah, Marselina Silvia dan Ernastuti (2010), Graduation Prediction Of Gunadarma University Students Using Algorithm And Naïve Bayes C4.5 Algorithm. Undergraduate Program, Faculty of Industrial Engineering. Gunadarma University.

[2] Berry, Michael J.A and Gordon S. Linoff (1993), Data Mining Techniques For Marketing, Sales, Customer, Relationship Management. Second Edition. Wiley Publishing, Inc. 2004.

[3] Quinlan, J. R. (1993), C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers.

[4] Kusrini, dan Emha Taufiq Luthfi (2009), Alogaritma Data Mining. Yogyakarta: Andi Publisher.

[5] David, Julie M. and Balakrishnan, Kannan (2010). Significance Of Classification Techniques In Prediction Of Learning Disabilities. International Journal of Artificial Intelligence & Applications (IJAIA), Vol.1, No.4, October.

[6] Kusrini dan Hartati, Sri. (2005). Implementation Of C4.5 Algorithm To Evaluate The Cancellation Possibility Of New Student Applicants At Stmik Amikom Yogyakarta. Gadjah Mada University, Mathematic and Natural Science Faculty, Yogyakarta, Indonesia.

[7] Larose, Daniel .T. (2005), Discovering Knowledge in Data. New Jersey: John Willey & Sons..

[8] Santosa, Budi (2007). Data Mining: Teknik Pemanfaatan Data untuk Keperluan Bisnis. Yogyakarta: Graha Ilmu.

[9] Suyanto (2011). Artificial Intelligence. Informatika. Bandung.

[10] Jiawei and K. Micheline (2008). Data Mining-Concepts and Techniques, Second Edition, Morgan Kaufmann - Elsevier Publishers, ISBN: 978-1-55860-901-3.

[11] Witten, Ian H. dan Eibe Frank (2005). Data Mining: Practical machine learning tools and techniques, Second Edition. Morgan Kaufmann. San Francisco.