pattern_recognation

by Anita Desiani

Submission date: 01-Jul-2022 09:28PM (UTC+0700)

Submission ID: 1865445816

File name: pattern_recognation_for_study.pdf (1.08M)

Word count: 3625

Character count: 18150

Journal of Physics: Conference Series

PAPER • OPEN ACCESS

Pattern recognition for study period of student in Mathematics Department with C4.5 algorithm data mining technique at the Faculty of Mathematics and Natural Science Universitas Sriwijaya

To cite this article: Sugandi Yahdin *et al* 2019 *J. Phys.: Conf. Ser.* **1282** 012014

View the article online for updates and enhancements.

Pattern recognation for study period of student in Mathematics Department with C4.5 algorithm data mining technique at the Faculty of Mathematics and Natural Science Universitas Sriwijaya

Sugandi Yahdin[1], Anita Desiani[1]*, Ali Amran[1], Desty Rodiah[2], Solehan[1]

[1]Mathematics Department of Universitas Sriwijaya, ,
[2]Informatics Department of Universitas Sriwijaya

*Corresponding author : anita_desiani@unsri.ac.id

Abstract : The study period pattern of student can facilitate the stakeholders to improve education services in their institutions. This study investigated the study period pattern of the graduated students from January 2014 until February 2018 with the C4.5 algorithm. The obtained data from three academic years 2011, 20112, and 2013 is 140 students with 18 frequently repeated subjects. The total attributes that used in the study are 19 attributes. These attributes contain 18 frequently repeated subject and one attribute for the label is from the study period of a student (punctual or not punctual). This study used the method of K-fold cross-validation with K = 10 and the percentage split with the percentage is 66% to measure the performance of the algorithm C4.5. The results of both methods showed that the subjects that affect sequentially in the study period of the student are a complex function, numerical methods, introduction to computer science, computer programs and basic physics. The accuracy of the results obtained from both the training process K-fold cross-validation amounted to 84.29% and 72.92% for percentage split. It shows that algorithm C4.5 is quite well in predicting the pattern of the study period of a student majoring in mathematics, Department of Universitas Sriwijaya.

1. Introduction

Educational Data Mining (EDM) is a trend that emerged in 2005, where data mining techniques are widely used to gain knowledge from the field of education [1]. The main objective of research in the field of EDM is to support decision making in educational institutions so that it can be useful for decision makers in the field of education. EDM has been widely used in a variety of research including the application of data mining for scholarship recommendations at Muhammadiyah Gubug High School [2], the classification of the graduation class of students of the Faculty of Communication and Informatics of Muhammadiyah University of Surakarta [3], the classification of student characteristics data [4], student classification programs drop out [5–7], and data mining applications use decision tree methods to display student final report [8].

One of methods used in data mining is C4.5 algorithm. The C4.5 algorithm is the best algorithm for making decision trees compared to other algorithms [9]. From several studies show that the C4.5 Algorithm method with decision tree techniques is preferred because it has advantages such as being able to process numerical and categorical data, producing rules that are easy to understand, because it is depicted in the form of images so that it can be seen which attributes affect and attributes which has no effect [1,2, 6, 9–12].

Mathematics Department The Faculty of Mathematics and Natural Sciences Universitas Sriwijaya is a study program that has a study load of at least 144 credits (semester credit units) or consists of 18 courses, consisting of 109 credits of compulsory courses and 35 credits of elective courses which can be taken in less than or equal to 8 regular semesters and less than or equal to 4 years [13]. Graduates' data for the period of January 2014 to February 2018 obtained at the Mathematics Department of the Faculty of Mathematics and Natural Sciences Universitas Sriwijaya (from the 3 student periods of 2011-2012 and 2013) were obtained at 8.63% of students who were able to complete their studies in less than or equal to 8 semesters. This shows that there are still many regular undergraduate students in the Mathematics Department who study more than 8 semesters. Study period is an important attribute for academic managers of the Mathematics Department. The department can minimize the failure to complete the study period of student by making plans, escorting studies, more intensive guidance and others.

One of many factors that can be an early identification for the determinant of the possibility of the study period and the level of student academic success is their success in taking courses in the early years of the first year of recovery. Students of the Mathematics and Natural Sciences Mathematics Department in the first and second year of college usually have to take courses that are required by the university, faculty and department. The overall compulsory subjects in the department are as many as compulsory courses must be compulsory for all students so that the success of students in lectures can be predicted early on from the success of students in the compulsory subjects they take. In the research will examine the pattern of study period (punctual or not punctual) of Mathematics Department students by using C4.5 data mining algorithm technique. This study is limited to the study time that can be completed by students (study period) who graduated in the graduation period from January 2014 to February 2018. The data taken came from three generations, namely 2011, 2012, 2013 with the curriculum used namely the 2007 and 2012 curriculum. The investigated pattern is the initial value obtained in the frequently repeated compulsory subject.

2. Literature Review

2.1. Classification Technique

Classification is a form of data mining techniques or methods that are included in the prediction category, which is a technique that can be used to predict or predict future data trends where testing utilizes a collection of classified data and attributes to determine additional outputs and classes[14]. One method in classification is c4.5 which is included in the decision tree method.

The concept of a decision tree is to change the data to be used as a decision tree model or better known as (decision tree) and then the results will be obtained in the form of rules[1]. The decision tree is an IF ... THEN set of rules that can be described in the form of a tree called a decision tree [15]. Each path in a tree is linked to a rule, which consists of a set of nodes, starting with the data at the root node, then selecting an attribute that meets the results of the branch child node (internal node), doing the recursive process on each leaf node (Leaf) is shown in Figure 1 as follows.

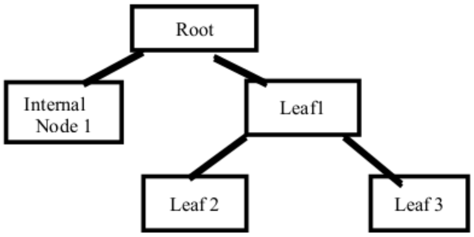

Figure 1. Decision Tree

2.2. Algorithm C4.5

The C4.5 algorithm uses the concept of information gain or entropy reduction to choose optimal branching. Suppose there is a variable x which has a number of k possible values with probabilities P_1, P_2, ..., P_k. Entropy describes the uniformity of data in variable *S*. Entropy *S* is calculated using equation1 [4].

Select attributes with roots, based on the results of the highest gain values of the existing attributes. To calculate the gain value the following formula is used [16]:

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{k} \frac{S_i}{S} \times Entropy(S_i) \qquad (1)$$

Remarks:
S : case set
A : attribute
n : number of partition attributes A
S_i : number of cases on the partition of -i
s : number of cases in S

So that the gain ratio results from the highest attributes. Gain is one of the test selection attributes that is used to select from the test attribute of each node on the tree. Attributes with the highest information gain are selected as a test attribute of a node. Calculation of looking for entropy values can be seen in equation 2 [17]:

$$Entropy(S) = \sum_{i=1}^{n} -p_i \times ln_2\, p_i \qquad (2)$$

Remarks:
S : case set
A : attribute
k : number of partition S
p_i : Proportion of S_i towards S

3. Methodology

3.1. Collecting Data

The data processed is data from Mathematics Department students of FMIPA UNSRI who have completed their studies starting from the graduation period in January 2014 to February 2018 coming from the 2011, 2012 and 2013 generations. The known data are the number of students who have completed his study from the 2011 class as many as 41 students, class of 2012 as many as 57 students and class of 2013 as many as 42 students.

The data obtained were 140 students who graduated during this period and 24 compulsory subjects namely introductory courses in computer science, calculus 1, general chemistry, general biology, calculus 2, mathematical logic, basic statistics, computer programs, discrete methods, calculus variables. , elementary linear algebra, financial mathematics, opportunity theory, numerical methods, ordinary differential equations, algebra 1, mathematical statistics, analysis 1, analysis 2, algebra 2, partial differential equations, insurance mathematics, basic physics, and complex functions they must take during the lecture period based on the applicable curriculum, the 2007 and 2012 curriculum.

3.2. Pre-Processing Data

Data selection or selection is done to choose which attributes will be used in the study. Of the 24 compulsory courses there are 18 compulsory subjects that are most often repeated by students, namely: Basic Physics, Multiple Variable Calculus, Calculus 1, Ordinary Differential Equations, Calculus2, Elementary Linear Algebra, Financial Mathematics, Discrete Methods, Analysis 1, Numerical Methods, Biology General, Complex Functions, Opportunities Theory, Mathematical Insurance, Partial Differential Equations, Mathematical Logic, Introduction to Computer Science, Computer Programs.

The total data used in the study is 140 data with 18 attributes of compulsory courses that are often the most repeated and 1 attribute of the study period of students who have graduated punctual or not punctual. The total attributes used in this study are 19 attributes.

3.3. Data mining process

Data of 140 students trained to be calculated according to existing calculations. The training process is done in 2 ways, namely:

1. Percentage Split

In this study the percentage chosen was 66% used as training data, and 34% of the remaining data was used as test data, which means that from the data 140 data were used as much as 48 data for training, and 24 data as test data.

2. K-fold Cross-validation

In this study, 10 K-fold values were chosen, meaning that from 140 students were divided into 10 groups, with 14 groups of students. The data used were 1 test data group and 9 training data groups which were conducted alternately.

4. Discussion

4.1. Percentage Split

At percentage split data is divided into two proportions, one proportion as test data and one proportion as training data. In this study the proportion of 66% was taken for the test data, and the remaining 34% for training data. From the test process, 34% data with percentage split obtained confusion matrix in table 4.9.

Table 1. Confusion matrix research data with percentage split

Correct Classification	Classification of	
	TT (Not On Time)	T (On Time)
	+	-
+	TP = 31	FN = 5
-	FP = 6	TN = 6

In the percentage split the first row of the first column with a value of 31 means that there are 31 data entered in the TT label classification that successfully predicted the class labeled TT, in the first row of the second column with a value of 5 means there are 5 data included in the TT label classification, but predicted as class labeled T. In the second row of the first column, which is worth 6, there are 6 data included in the T label classification, but it is predicted as a class labeled TT. In the second row of the second column which is worth 6 means that there are 6 data included in the classification of the T label that has successfully predicted the class labeled T. From table 4.9 can be calculated the value of accuracy, precision and recall for each class. Accuracy value was obtained at 84.28%. The recall value for each class is, for the class not on time (TT) the recall value is 92.67%, while for the Punctual class (T) the recall value is 40.63%. The precison values for each class are as follows for the class of Not Punctual (TT) of 88.39%, and for the class on time (T) of 67.85%.

4.2. 10-foldCross-Validation

From the C.5 algorithm test process with 10-cross-validation, confusion matrix is obtained in table 2.

Tabel 2. Confusion matrix research data with 10-fold cross-validation

Correct Classification	Classification of	
	TT (Not Punctual)	T (Punctual)
	+	-
+	TP = 99	FN = 9
-	FP = 13	TN = 19

From table 2, In the first row and the first column is 99 means that there are 99 data included in the TT label classification that is successfully predicted the class labeled TT, in the first row of the second column which is 9 means there are 9 data included in the TT label classification, but predicted as class labeled T. In the second row of the first column, which is 13, there are 13 data included in the T label classification, but it is predicted as a class labeled TT. In the second row of the second column which is worth 19 means there are 19 data included in the classification of the T label that successfully predicted the class labeled T. From the confusion matrix on k-fold cross validation obtained accuracy of 84.29%, recall for TT class of 91.7 % and for class T is 40.63%, prescision for TT class is 88.39% and for class T is 67.85%.

Results analysis by referring to the decision tree image can be seen in table 3:

Table 3. Results of comparison of Cross-Validation and Percentage Split performance

Method C4.5 algorithm	**Test Method**			
	Percentage Split (%)		Cross-Validation (%)	
	T	TT	T	TT
Accuracy	72.92		84.28	
Precision	54.54	83.74	67.85	88.,39
Recall	50	86.11	40.63	91.67

From table 3, The accuracy of the percentage split is only 72.92%, although above 50% the accuracy of the model is not good enough. The training and test method is carried out with the cross validation method showing a better accuracy of 84.28%, which means that the prediction model produced is good enough to be used. The precision value and recall value for the data for TT class is greater than T. This is due to the amount of TT class data more than T means that the number of students who cannot complete their study period is far more than students who can complete their studies on time.

The results of the C4.5 algorithm processing obtained by the compulsory subject patterns that affect the student study period are as follows :

- **IF** the subject of Complex Functions is worth B, C, and D, **THEN** it is predicted that the period of study is Not Punctual **(OR) IF** the subject of Complex Functions is A and the numerical method is A, **THEN** it is predicted that the study is punctual **(OR) IF** the function is worth A and the method numerical value of C or D **THEN** it is predicted that the period of study is not punctual, **OR IF** the function of the complex is worth A and the numerical method is worth B and Introduction to computer science is A, **THEN** the study period is punctual **OR IF** the function of the complex is A and the numerical method is B and Introduction computer science is worth B and the discrete method is worth A **THEN** it is predicted that the study period will be punctual.
- **IF** the course Complex functions are worth A and the numerical method is B and the introduction of computer science is B and the discrete method is C and D **THEN** it is predictable that the period of study is not Punctual **OR IF** the courses Complex functions are A and numerical methods are B and Introduction to computer science value of C and **IF** the computer program is worth A, C and D **THEN** it is predicted that the study period is not punctual.
- **IF** the subject of Complex Functions is worth A and the numerical method is B and Introduction to computer science is C and the computer program is B, **THEN** the study period is Punctual **OR IF** the Complex Functions are A and numerical methods are B and Introduction to computer science is D, **THEN** it is predicted that the study period was not punctual.
- **IF** the course Complex functions are worth A and the numerical method is B and the introduction of computer science is B and the discrete method is worth B and the basic physics is A (OR) B (OR) C **THEN** it is predicted that the study will be punctual.

- **IF** the function of the complex is worth A and the numerical method is B and the introduction of computer science is B and the discrete method is worth B and the basic physics is D **THEN** it is predicted that the study will not be punctual.

5. Conclusion

The most influential compulsory courses in the student study period are complex functions, numerical methods, introduction to computer science, discrete methods, computer programs, and basic physics by looking at the influence of each of the courses obtained. The test results show that using the C4.5 algorithm is quite good in predicting the study period of students based on the compulsory subjects they take, although the accuracy obtained by using percetage split is only 72.92% but the accuracy of the model used can be higher K-fold Cross-Validation, which is 84.28 %.

References

[1] Kusrini and Luthhfi E T 2009 *Algoritma Data Mining*, 1st ed. Yogyakarta: Andi.

[2] Dina M 2014 Penerapan Data Mining Untuk Rekomendasi Beasiswa Pada SMA Muhammadiyah Gubug Menggunakan Algoritma C4.5, in *Seminar Nasional Aplikasi Sains & Teknologi (SNAST).*

[3] Nugroho Y S 2015 Klasifikasi dan klastering mahasiswa informatika universitas muhammadiyah surakarta, *Univ. Res. Colloq. 2015,* 89–98.

[4] Rahmayuni I 2014 Perbandingan performansi algoritma c4.5 dan cart dalam klasifiksi data nilai mahasiswa prodi teknik komputer politeknik negeri padang, *J. Teknoin*, **2** (1), 40–46.

[5] Anis A 2012 Penerapan Algoritma C4.5 Pada Program Klasifikasi Mahasiswa Dropout, in *Seminar Nasional Matematika*

[6] Dekker G W, Pechenizkiy M and Vleeshouwers J M 2009 Predicting Students Drop Out : A Case Study, in *The 2ndInternational Conference on Educational Data mining* 41–50.

[7] Jun J 2005 *Dropout of adult learners in e-learning*, The University of Georgia.

[8] Novianti T and Aziz A 2015 Aplikasi Data Mining Menggunakan Metode decision Tree untuk menampilkan laporan Hasil Nilai Akhir Mahasiswa Fakultas Teknik UMSURABAYA , *Netw. Eng. Res. Oper. [NERO]*, **1** (3).

[9] Hssina B, Merbouha A, Ezzikouri H and Erritali M 2014 A comparative study of decision tree ID3 and C4 . 5, *Int. J. Adv. Comput. Sci. Appl. Spec. Issue Adv. Veh. Ad Hoc Netw. Appl.* **4**(2) 13–19.

[10] Han J and Kamber M 2006 *Data Mining : Concepts and Techniques*, 3rd ed. USA: Morgan Kauffman.

[11] Yadav S, Bharadwaj B and Pal S 2012 Data mining applications: A comparative study for predicting student's performance, *Int. J. Innov. Technol. Creat. Eng.* **1** (12) , 13–19.

[12] Romero C and Ventura S 2013 Data mining in Education, *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **3** 12–27.

[13] Unsri J M F 2013 *Panduan Kurikulum Program Studi Matematika Tahun Ajaran 2014/2015*.

[14] Fayyad U 1996 *Advances in Knowledge and Data Mining*. American Association for Artificial Intelligence Menlo Park.

[15] Jiawei V, Micheline K and Jian P 2011 *Data Mining: Concepts and Techniques*. San Francisco, CA, USA: Elsevier.

[16] Mantas C J and Abellán J 2014 *Expert Systems with Applications Credal-C4 . 5 : Decision tree based on imprecise probabilities to classify noisy data*, **41** (10) Elsevier Ltd.

[17] Hartanto D and Seng H 2014 Implementasi Data Mining dengan Algoritma C4.5 untuk Memprediksi Tingkat Kelulusan Mahasiswa *ULTIMATICS*, **4** (1).

pattern_recognation

ORIGINALITY REPORT

18%	15%	16%	9%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

MATCH ALL SOURCES (ONLY SELECTED SOURCE PRINTED)

6%

★ repository.unja.ac.id

Internet Source

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		