Tuition Single Classification using Decision Tree Method and C4.5

Baihaqi Siregar¹, Erna Budhiarti Nababan¹, Noviyanti Sagala¹, Uli Andayani¹ and Fahmi²

¹Department of Information Technology, Faculty of Computer Science and Information Technology
²Department of Electrical Engineering, Faculty of Engineering University of Sumatera Utara, Medan, Indonesia

Abstract. Tuition Single (UKT in Indonesia) is an operational cost measurement standard for universities in Indonesia which is applied by the government since 2013. This standard is enforced for every public university students who pass entrance university test. The problem with this standard is still found that there are students with tuition single category which does not fit with their condition, thus resulting in the incriminating of students with difficult economic situations. The usage of the decision tree and C4.5 algorithm for classification is able to process continuous data and build model efficiently with a large number of data. The input variables for tuition single classification are the income of the financier, the status of the children within the family, the taxable value of land and building, and the ownership of the vehicle. This work uses a multiclass variable. The output variables are group 1 to group 7. The accuracy of the system is 80.25% where the groups with the most students are group 3 and group 7. The decision tree built is also quite big considering the number of the variables and the variation of data used in each variable.

1. Introduction

Data mining has attracted lot of attention in the research industry and in society as a whole in recent years, due to enormous availability of large amount of data and the need for turning such data into useful information and knowledge. Data mining is utilized by universities to handle student data, such as tuition waiver. Currently, the tuition waiver system for students who pass the entrance university test in all public universities in Indonesia is called UKT (Tuition Single).

Tuition single consist of 2 types i.e. full and equitable tuition single. In University of Sumatera Utara, Indonesia, the equitable tuition single is college financing system in the range of category I to category VII (highest) per semester in which the nominal of tuition single is determined by the economy ability of financier based on various aspects whereas full tuition single is college financing system in the category VII (the highest 1 category) per semester. Nominal tuition single depends on the department where the students received. Tuition waiver becomes a serious problem in public universities now. This impact to students who got equitable tuition single system. According to the evaluation of universities, they found students with equitable tuition single category that do not correspond to the economic capabilities of the students.
Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. In this work, a classification system was developed using decision tree which can produce a model with rules that are human-readable and interpretable and C4.5 classifier which is the most popular and powerful decision tree classifier is able to process continuous data and to handle missing values. Decision trees are built through attributes selection process which based on the highest value of the entropy and gain of the existing attributes, then create a child node for each value and divide the selected attributes into the branches until the elected attribute has the same class. In this paper, section 2 describes related work to decision tree C4.5 and classification. Section 3 describes the methodology. Section 4 shows the results and discussion. Finally, section 5 concludes the paper and outlines further work.

2. Related Work
There is no previous research which has been used this data so the related work only focus on the method. Decision tree C4.5 has been implemented in order to predict the diabetes disease. The performance of the system shows that the system has great potential up to 81.27 correctly. Recently, data mining often implemented in educational data to predict the performance of the student. This work shows that students past academic performance can be used to create the model using decision tree algorithm and the model is successfully identified the students who are likely to fail. Classification technique can be implemented in breast cancer datasets. This study used a hybrid approach, CART classifier with feature selection and bagging technique in order to enhance the classifier accuracy. A hybrid approach has also been implemented by researchers in case of determination of major specialization. Fuzzy Mamdani and the C4.5 algorithm are used to analyze the determination of major specialization in informatics engineering courses of Universities Raya Serang. Though various classification techniques are widely used for Disease Prediction, Decision Tree classifier is selected for its simplicity and accuracy.

In this work, the proposed usage of the decision tree and C4.5 algorithms is to classify tuition single using income, the status of children within the family, the taxable value of the land and building, and the ownership of vehicles variable.

3. Methodology
The general architecture of proposed method in this work is shown in Figure 1. The steps in classifying tuition single in this work are as follows:

Preprocessing data mining includes data cleaning on training and testing dataset; Data selection and data transformation are performed on training data set to get attribute and valid data; attributes are formed through the calculation of GINI; calculate the entropy value of each instance and the gain of each attribute to form a decision tree; determine the class label; the establishment of rule of trees; testing data is processed using rules that have been formed. Then, it will produce a nominal category and the category of tuition single. Each of the steps being taken is described as follows:

3.1. Dataset
Data used in this research is student data of the University of Sumatera Utara in 2013. The total data is 6063. Then, data is divided into 2 datasets which will be used to train the decision tree C4.5 and to obtain the accuracy of the method. The composition of training data and testing data is 80% and 20%. These data provide a real picture of how a student classified into a particular category based on the economic capabilities.
3.2. Pre-process

Further explanation of a series of steps on the general architecture of the proposed method is as follows:

1) Data Cleaning

There are 2 ways to handle missing values which are reducing the data set and the most common value of an attribute restricted to a concept. The reduction of the data set is performed by the elimination of samples (rows) with missing values or elimination of attributes with missing values. This technique is performed on 37 rows. The total data after the cleaning process is 6,026 rows.

2) Data Selection

Data selection is performed in order to analyze the trend of the data thus obtained the attributes in determining the group of tuition single. The result of this process is the income of financier, the status of the student within the family, the ownership of vehicles and taxable value of land and building.

3) Data Transformation

The data sets are transformed into a form which is more suitable for mining. The datasets are separated and categorized and so as the dataset will be easier for mining and obtained knowledge discovery from the datasets.

3.3. Decision Tree and C4.5 algorithm

We begin by specifying the notation that will be used in the rest of this paper. We use \( A \) to denote attributes, \( n \) to denote number of attributes of partition \( A \), \( pi \) to denote the proportion of \( S \), \( S \) to denote the set of cases, \( |S_i| \) to denote the number of ith partition, \( |S| \) to denote the number of cases in \( S \), \( p(j|t) \) to denote relative frequency class \( j \) on node \( t \), \( ni \) to denote the amount of data of child node, and \( t \) to denote the amount of data of node \( p \).

1) Attributes Formation

The formation of the attributes is performed using GINI formula. The following steps are to form attributes:
• Sort the values of each variable (ascending order) as shown in Table 1.

| NO. | INCOME   |
|-----|----------|
| 1   | 1200000  |
| 2   | 1213900  |
| 3   | 1500000  |
| 4   | 1850000  |
| 5   | 2100000  |

• Calculate the average of the two adjacent values in each variable as shown in Table 2.

| NO. | INCOME   | AVERAGE VALUE |
|-----|----------|---------------|
| 1   | 1200000  | 1206950       |
| 2   | 1213900  | 1356950       |
| 3   | 1500000  | 1500000       |
| 4   | 1850000  | 1675000       |
| 5   | 2100000  | 1975000       |

• Calculate the GINI using equation (1) and (2)

\[
GINI (\tau) = 1 - \sum_j [p(j|\tau)]^2
\]  

\[
Gini \ Split = \frac{\sum_{i=1}^{k} n_i}{\tau} GINI (i)
\]  

GINI calculation performs for all average value. The calculation result of income variable is shown in Table 3.

| AVERAGE VALUE | 1206950 |
|--------------|---------|
| Separator    | <=      | >       |
| Group 1      | 392     | 26      |
| Group 2      | 740     | 3       |
| Group 3      | 1008    | 96      |
| Group 4      | 412     | 144     |
| Group 5      | 270     | 182     |
| Group 6      | 192     | 203     |
| Group 7      | 131     | 1026    |
| GINI         | 0.7244  |
- Determine the split of each variable
  The smallest value of the average GINI in each variable would be the split for that variable. List of the split is shown in Table 4.

![Table 4. List of Splits](image)

| SEPARATOR                      | <= 51594600 | > 51594600 |
|--------------------------------|-------------|------------|
| Income                         |             |            |
| The status of the child within the family | 1          | 1          |
| The taxable value of the land   | 14000       | 14000      |
| The taxable value of the building | 365000    | 365000     |
| Number of cars                  | 0           | 0          |
| Number of motorcycles           | 1           | 1          |

2) Calculate Entropy and Gain
The goal of calculating entropy and gain is to form a decision tree. Formation of the decision tree essentially has the early stages of determining the root of the decision tree. The step of determining the root is as follows:

a. Calculate the total case of each group for each attribute
b. Calculate the entropy total value using equation (3).
\[
Entropy(A) = \sum_{i=1}^{n} - pi \times \log_2 pi
\]  
(c) Compute the entropy value of each attribute
(d) Compute the gain value of each variable using equation (4)
\[
Gain(S, A) = Entropy(S) - \sum_{i=1}^{m} \frac{|S_i|}{|S|} \times Entropy(S_i)
\]

To determine the root node of the decision tree is based on the highest gain value. In this work, the root node is income variable. The attributes of income variable are <= 51,594,600 and >51,594,600 which used as the split.

3.4. Knowledge Representation
Knowledge obtained from the decision tree then represented in the form of IF-THEN rules.
The value of an attribute is antecedent part (IF), a leaf of a decision tree will be consequent part (THEN). This rules to be very helpful in understanding the classification model, especially if the size of the decision tree is oversize. The following rules described a sample of each group.

Group 1: If ‘income’ <= 51,5946K && ‘the status of the child’ <= 1 && ‘the taxable value of the land’ > 14000 && ‘the taxable value of building’ <= 365000 && ‘number of car’ > 0 && ‘number of motorcycle’ > 1
Group 2: If ‘income’ <= 51,5946K && ‘the status of the child’ > 1 && ‘the taxable value of the land’ <= 14000 && ‘the taxable value of the building’ <= 365000 && ‘number of car’ <= 0 && ‘number of motorcycle’ <= 1
Group 3: If ‘income’ <= 51,594.6K && ‘the status of the child’ > 1 && ‘the taxable value of the land’ <= 14000 && ‘the taxable value of the building’ > 365000 && ‘number of car’ <= 0 && ‘number of motorcycle’ <= 1

Group 4: If ‘income’ <= 51,594K && ‘the status of the child’ > 1 && ‘the taxable value of the land’ <= 14000 && ‘the taxable value of the building’ <= 365000 && ‘number of car’ > 0 && ‘number of motorcycle’ > 1

Group 5: If ‘income’ <= 51,594.6K && ‘the status of the child’ > 1 && ‘the taxable value of the land’ > 14000 && ‘the taxable value of the building’ <= 365000 && ‘number of car’ > 0 && ‘number of motorcycle’ > 1

Group 6: If ‘income’ <= 51,594.6K && ‘the status of the child’ <= 1 && ‘the taxable value of the land’ <= 14000 && ‘the taxable value of the building’ > 365000 && ‘number of car’ > 0 && ‘number of motorcycle’ <= 1

Group 7: If ‘income’ > 51,594.6K && ‘the status of the child’ <= 1 && ‘the taxable value of the land’ <= 14000 && ‘the taxable value of the building’ <= 365000 && ‘number of car’ <= 0 && ‘number of motorcycle’ <= 1

3.5. Pattern Evaluation
Based on the rules obtained, group 3 and group 7 have the highest number of rules, this can happen because the amount of the data of these groups are larger than the other groups. The more data each group were used, the more numerous and varied rules generated by the group. Moreover, the class label is given at the last node indicating that the data in each group have not reached zero yet, so it requires further elaboration.

3.6. Determine Classification Result
The formation of the node will continue to be done and will stop if it meets one of the following conditions:

- Only one group subset of the sample left, then label the class is the group.
- The built decision tree has already used all the variables but there is more than one group left, the label given to the group class with the largest number of students. In this study, the determination of the class label fulfills condition b, thus, produce large decision tree.

4. Result And Discussion
System testing is performed to ensure that the elements or components of the system have been functioning as expected and to check the performance of the algorithm. The total of the testing dataset is 1201 data. The experimental result is in Figure 2.
Based on figure 2, the largest number of student is in group 7 because the amount of training and testing data sets for group 7 is the largest thus the resulting rules are more varied. Otherwise, the total student for other groups is small thus the resulting rules are slightly varied.

5. Conclusion
The accuracy of implementation of decision tree C4.5 to classify the student was 80.52%. The more class label and the more data used, the more numerous and varied the obtained rules. In this work, the training process time was about 7 minutes because the attribute formation process needed a long time. Future works involve further enhancement of the method and implementation of future selection in order to obtain the best parameter for classifying the student.

References
[1] P. Kaur, M. Singh, G.S. Josan, Classification and prediction based data mining algorithms to predict slow learners in education sector. International Conference on Recent Trends in Computing (ICRTC), 57 (2015) 500-508.
[2] Regulation of the Minister of Education and Culture of the Republic of Indonesia Number 55 Year 2013 about the Tuition Fee and Tuition Single on College of Environmental Affairs in the Ministry of Education and Culture.
[3] B. Baradwaj, S. Pal, Mining educational data to analyze students’ performance. International Journal of Advanced Computer Science and Applications (IJACSA), 2 (2011).
[4] Q. A. Al-Radaideh, A. N. Eman, Using data mining techniques to build a classification model for predicting employee performance. International Journal of Advanced Computer Science and Applications (IJACSA). 3 (2012) 144-151.
[5] Purushottam, S. Kanak, S. Richa, Diabetes mellitus prediction system evaluation using C4.5 rules and partial tree. International Conference of Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), (2015) 1-6.
[6] R. Kabra, R. S. Bichkar, Performance prediction of engineering students using decision trees. International Journal of Computer Applications, 36 (2011).
[7] D. Lavanya, K. U. Rani, Ensemble decision tree classifier for breast cancer data. International Journal of Information Technology Convergence and Services, 2 (2012).
[8] Harsiti, T. A. Munandar, H. T. Sigit, Implementation of fuzzy-C4.5 classification as a decision support for student choice of major specialization. International Journal of Engineering Research & Technology. 2 (2013).
[9] K. Thenmozhi, P. Deepika, Heart disease prediction using classification with different decision tree techniques. International Journal of Engineering Research and General Science. 2 (2014).
[10] M. Kantardzic, Dealing With Missing Values. Data Journal of Systems Integration, 51 (2014) 65 - 176.