Name ID

Lab-Decision-Tree

DAKDL University Case Study

DAKDL managers noticed that student major selection is very important factor of his/her success. So, they decided to discover useful knowledge in order to help students selecting appropriate major. They want to find out the *characteristics* of good students for each major. At DAKDL, there are 6 majors of engineering, which are computer, electrical, mechanical, chemical, environmental and civil. Good students (GPA >= 3.0) are those that obtain good *grade point average* (GPA) for their graduation in a given major.

More precisely, DAKDL managers want to determine which majors should be appropriate to which categories of students. *Characteristics* of students are based on all combinations of attributes such as student's *profile*: age during first year study, gender, region, etc. and student's *university performance*: grade obtained in each course in the first year. Several methods are possible to reach the objective. A simple method is to construct a classification model to predict the most appropriate major for a given student. Decision tree can be used for classification model construction.

Your task is to do necessary steps for pre-processing data and to discover useful knowledge. Following is example of knowledge is the form of rule that could be derived from decision tree model:

If (417168 is Low) and if (403111 is High) and if (204111 is Low) and if (420112 is Low) and if (Gender is M) then Class –Computer Engineering

Discovered knowledge can help students to select the appropriate major (among 6 possible majors) according to their characteristics when they are to enter in the second year.

Name ID

First analysis of your data:

Q1 What is the number of good students who have already graduated? Good students are those, which graduated with GPA greater or equal to 3.0.

Q2 What is the number of bad students who have already graduated? Bad students are those, which graduated with GPA less than 3.0.

Q3 For each department, give the number of students who have already graduated

Department Name	Number of occurrences	Percentage
Computer_Engineering		
Civil_Engineering		
Electrical_Engineering		
Chemical_Engineering		
Mechanical_Engineering		
Environ_Engineering		

Q4 What is the total number of first year students?

Data Pre-processing:

Q5 What are all necessary tables to be used in this data-mining project?

Pre-process your data in order to obtain the table to be input for the decision tree construction technique. We suggest the following techniques:

1. Discretize Age at the first year study using the following range:
   - Low = 16,17,18
   - Medium = 19,20,21
   - High = 22,23
2. Discretize grade for each course of the first year study using the following range:
   - Low = 0.0,1.0,1.5
   - Medium = 2.0,2.5
   - High = 3.0, 3.5,4.0

Training Phase:

Build a model in order to predict which majors should be appropriate to which students using decision tree. Training data consist of good students that have already graduated. Recall that good students are those, which graduated with GPA greater or equal to 3.0. Following is the list of attributes necessarily for construction data classification model:

1. Gender
2. Age at the first year study
3. Region
4. Department
5. Grade obtained for each course during the first year study (204111,204222,204333,204444,204555,204666)

- ❑ Following is data format and example of data to be input for the decision tree construction technique

Stu_ID	Gender	Age_1_year	Region	Dept	204111	204222	204333	204444	204555	204666
37058063	male	Medium	Central	Civil-Eng	Medium	Low	Low	Medium	High	Low
37058167	male	High	South	Electrical-Eng	High	High	Medium	High	Medium	Low
…………	….				…..	…..				

Q6 Using Decision tree

Q6.1 What is the class label attribute? How many classes to be predicted?

Q6.2 Split data into training set (70%, stratified sampling) and test set (30%, stratified sampling). What is the number of training records?

Q6.4 What is the accuracy of the model with the hold-out method? Using the following parameters: Criterion: gini_index, minimal size for split: 55, minimum leaf size: 50

Q6.5 How many levels are in your decision tree?

Q6.6 Explain your tree in terms of predictors to predict Department

Q6.7 Which attribute(s) is the most related to Department?

Q6.8 Give precision and recall for each class?

Department Name	Precision	Recall
Computer_Engineering		
Civil_Engineering		
Electrical_Engineering		
Chemical_Engineering		
Mechanical_Engineering		
Environ_Engineering		

Q6.9 Which major(s) is the most accurate? Why?

Model Improvement:

Given the fact that the number of students is different in each department, and then the obtained model cannot be used to predict correctly the best major to be selected since classification error would be very high. If electrical department can accept more than 50% of all engineering students, then the model will predict that almost every student should select electrical as his/her major. Here, the objective is to improve the previous classification model.

Possible model improvement consists in building a *classification model for each major*. A decision tree for each major will have many branches and levels, but at the leaves there will be two classes: *good* and *bad* students. Good students are *those graduated with GPA >= 3.0*, and bad students are those *graduated with GPA < 3.0* in a given department. At the training step, input data are separated into two groups: good and bad students in a given department.

Model improvement-training phase

Q9 Build a decision tree model for the department of **Electrical Engineering**

Q9.1 What is the class label attribute? How many classes to be predicted?

Q9.2 What is the accuracy of the model 10-fold cross-validation method?

Q9.3 What is the number of training records to be input for the decision tree construction technique?

Q9.4 Explain your tree in terms of predictors and the target class. How many levels are in your decision tree?

Q9.5 Which attribute(s) is the most related to Electrical Engineering?

Q6.6 Give precision and recall for each class

Name ID

Application phase

Q10 Apply your training model in order to predict whether a new student is going to good or bad student in electrical engineering.

Student_ID		
5342		
3982		
5941		
5942		
5662		