

INFS 755 (Fall 2008)

Assignment 2 (Due on 10/13/2008) (Datawarehousing, Classification) by Huzefa Rangwala

This is an individual assignment. Please ensure that assignment is submitted in class in hard copy before start class. No late submissions allowed. The first part of the assignment introduces you to reading a research article. It is on data warehousing and serves as a complementary reading to the book. The second part of the assignment are a few questions on classification algorithms (small exercise in weka) from Chapter 4.

Part 1 (50 points)

Read the research article in the business intelligence journal (<http://www.tdwi.org/research/display.aspx?ID=6810>). Based on your reading of this article do the following.

1. What are the differences between the data warehouse and operational databases from a user perspective. (5 points)
2. What does Kimball think about snowflake schema? Do you agree with his point of view? (10 points)
3. Understand the steps involved in the converting the ER model for Retail Sales System to a dimensional model. Now design a data-warehouse for George Mason University consisting of the following dimensions: student, course, semester, and instructor. One set of measure I had like to see is student GPA. Design your dimensions in such a way that you could answer all the sub-questions below. Please read all questions before beginning your design.
 - a. Draw a star-schema model
 - b. Discuss what are the benefits of your designed data warehouse. What kind of information can I extend quickly and efficiently?
 - c. Extend the star schema to snow flake schema. Ensure that you came up with attributes of dimension tables in part (i) that allowed you to do this.
 - d. What OLAP operation will allow you to extract the following information from your warehouse (ex: drilling, slicing?)
 - i. Toughmost grader among the students in the Volneague School of IT &E.
 - ii. The average GPA for the girl students in the Computer Science Department for the year 2008. (30 points)
4. Based on the reading how are ER modeling and dimensional modeling related? (5 points)

Part 2 (50 points)

1. Given the following training examples for the target concept EnjoySport . (20 points)

Example	Sky	AirTemp	Humidity	Wind	Water	Forecast	EnjoySport
1	Sunny	Warm	Normal	Strong	Warm	Same	Yes
2	Sunny	Warm	High	Strong	Warm	Same	Yes
3	Rainy	Cold	High	Strong	Cool	Change	No
4	Sunny	Warm	High	Strong	Cool	Change	Yes
5	Sunny	Warm	Normal	Weak	Warm	Same	No

Draw the decision tree that will be learnt by ID3 based on the training examples. Show the value for information gain for each of the candidate attributes for each step in the growing tree.

Now load the dataset in weka and learn a decision-tree using the id3, SimpleCART, and J48 classifiers. Remember to click on “Use training set”. Compare the three trees you generated. Which one of them you think has the best performance on an unseen test set and why ?

2. Consider the following data set for a binary class problem (below) (10 points)

- Calculate the information gain when splitting on A and B. Which attribute would the decision tree induction algorithm choose?
- Calculate the gain in the Gini index when splitting on A and B. Which attribute would the decision tree induction algorithm choose?

A	B	Class Label
T	F	+
T	T	+
T	T	+
T	F	-
T	T	+
F	F	-
F	F	-
F	F	-
T	T	=
T	F	-

3. If you had to choose between the naïve Bayes and k-nearest neighbor classifiers, which would you prefer for a classification problem where there are numerous missing values in the training and test data sets? Indicate your choice of classifier and briefly explain why the other one may not work so well? (10 points)

4. Explain why accuracy can be misleading in measuring the performance of classification models. Give an example. (5 points)

5. How do you handle missing valued data while learning decision trees ? (5 points).