Term Project: Classification on Mars Crater Dataset

Assigned Date: Wednesday, October 20, 2010

Educational Goal

Become familiar with advanced data structures and algorithms using real-world Mars data.

Phase I: Classification using Weka API

(200 points)

Due: 4:00 PM Monday November 15, 2010

Requirements

- **Dataset**: The data set is in CSV format (Comma-Separated Values). Download the file from http://www.cs.umb.edu/~ding/out/cs310_data/train.csv (~5.1MB).

  Note that you should not open the CSV file directly using MS Excel Spreadsheet because the data would be crashed if you do so. Read the file from a Java program.

  Data set description:

  Each crater candidate has 1089 attributes (Columns = 1 to 1089). Column 1090 is the class label, where 1 is for crater and 0 is for non-crater.

- Convert those data sets from CSV format to Weka ARFF format using a Java program.
- **Classification:** Use Weka J48 to build a classifier using 60% of the data as the training set and 40% of the data as the test set. Weka Java API provides correspondent parameter settings for training set and classifier parameters. You may use default parameters suggested by Weka.
- Report accuracy of the classification, number of correctly and incorrectly classified examples. Save correctly classified examples into correct.csv file and misclassified examples into wrong.csv file.

Submission Requirements

1. Write a report no less than 200 words to discuss your choice of data structures and algorithm design.
2. Prepare a readme file for your TA to run your project on his machine.

3. Generate Javadoc of your project. Your program should be well-documented. Variable names and function names should be self-descriptive. Major functions should be explained clearly The program outputs should be clearly presented.
4. Test your program thoroughly. Submit the outputs of your program.
5. Zip all the files. One submission per team. Save the file as CS310_ teamNumber. For example, Team 1 should name their file as *CS310_team1.zip.* Turn in the paper copy and soft copy of the assignment. Submit the softcopy of the file through your UMassOnline account at http://boston.umassonline.net/index.cfm. Submit the paper copy along with the cover page in class. Paper copy should be bound firmly together as one pack (for example, staple, but not limited to, at the left corner). 5 points will be deducted for unbounded homework.
6. The softcopy should include all the programs, readme file, Javadoc file, program outputs, and correct.csv and wrong.csv. The paper copy should include all the files of the softcopy, excluding correct.csv and wrong.csv.
7. No hard copies or soft copies results in 0 points.