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 II: Feature Selection Using Randomized Algorithms

(300 points)

Due: 5:30 PM Monday December 13, 2010

Requirements

- **Datasets**: The data set is in CSV format (Comma-Separated Values). Download the file from:

  training data set: http://www.cs.umb.edu/~ding/out/cs310_data/train.csv (~5.1MB; same data set used in Phase I).

  Test data set: http://www.cs.umb.edu/~ding/out/cs310_data/TestSet3.csv (~21MB).

- **Classification:** Use Weka J48 to build a classifier using training data set train.csv.
- **Feature Selection:** Design a randomized algorithm to find the best feature subset out of the total 1089 features that can achieve the highest F1 score on test set TestSet3.csv using training set train.csv. The following link explains how to calculate F1 score (F-measure): http://www.cs.umb.edu/~ding/classes/470_670/notes/evaluation_matrix.pdf. The instructor will explain the evaluation matrix in class on Wednesday November 17, 2010.
- **Randomized Algorithm:** What is the time complexity of the algorithm? Prove it. Draw a flowchart of your randomized algorithm.
- **Data Structures:** What is the major data structures used in your program? Justify your design.
- Write an experiment report to discuss your experimental results, including detailed parameter settings and experimental results. Discuss your experiences in writing these programs. What was the most difficult part? What development tools (IDE, etc.) did you use? Did you develop on UNIX, Windows or any other platforms? Did you have any problem recompiling and running Java with Weka? If yes, how did you fix it?
- **Note : Please do all the tasks of classification, feature selection and randomized algorithm in ONE program. Demonstrate all the functions of the project in one run. Do not divide those tasks into separate programs.**

Submission Requirements

1. Prepare a readme file for your TA to run your project on his machine.
2. 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.
3. Test your program thoroughly. Submit the outputs of your program.
4. 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.
5. The softcopy should include all the programs, readme file, Javadoc file, program outputs, and reports. The paper copy should include all the files of the softcopy.
6. No hard copies or soft copies results in 0 points.