Term Project: Feature Selection on Mars Crater Dataset

Assigned Date: Wednesday, October 20, 2010

Educational Goal

Apply AI techniques to real-world Mars data.

Phase I: Uninformed Search on Feature Selection

(200 points)

Due: 4:00 PM Monday November 15, 2010

Requirements

- **Training set and Test sets**: The training set and 3 test sets are in CSV format (Comma-Separated Values) . Download the files from http://www.cs.umb.edu/~ding/out/cs310_data/data.zip (~56 MB)

  Note that you should not open the CSV files directly using MS Excel Spreadsheet because the data would be crashed if you do so. Read the files using a programming language, for example, Java or Matlab.

  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. Implement a script for the data conversion, using a language of your choice.
- **Supervised Learning:** Use an **uninformed search algorithm of your choice** to find the best feature subset out of the total 1089 features that can achieve the highest accuracy on the training set using 10-fold cross validation.

  In 10-fold cross-validation, the original sample is randomly partitioned into 10 subsamples. Of the 10 subsamples, a single subsample is retained as the validation data for testing the model, and the remaining 9 subsamples are used as training data. The cross-validation process is then repeated 10 times (the folds), with each of the 10 subsamples used exactly once as the validation data. The 10 results from the folds then can be averaged (or otherwise combined) to produce a single estimation. The advantage of this method over repeated random sub-sampling is that all observations are used for both training and validation, and each observation is used

for validation exactly once. 10-fold cross-validation is commonly used in K-fold cross-validation (K can be any integer > 2).

- Use Weka LibSVM to build a classifier from the training set and use the built classifier to classify the 3 test sets, using the feature subsets you have selected. Weka Java API provides correspondent parameter settings for cross validation and classifier model, etc.

Submission Requirements

1. Write an experiment report to discuss your experimental results, including detailed parameter settings and experimental results. Submit the paper copy of the report, and source code of the scripts 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.
2. Submit the softcopy of the report and scripts through your UMassOnline account at http://boston.umassonline.net/index.cfm.
3. Zip all the files. One submission per team. Save the file as sdm_ teamNumber. For example, Team 1 should name their file as *sdm_team1.zip.*
4. No hard copies or soft copies results in 0 points.