Term Project: Detection of Small Craters in High Resolution Planetary Images Using Shape and Texture Features

Educational Goal

Apply spatial data mining techniques to real-world applications.

Phase II: Active Class Selection

(100 points)

Due: 8:30 PM Thursday April 29, 2010

Requirements

- Active Class Selection (ACS) addresses the question: if one can collect n additional training instances, how should they be distributed with respect to class? Active Class Selection is a new class of problems for multi-class supervised learning. If one can control the classes from which training data is generated, utilizing feedback during learning to guide the generation of new training data will yield better performance than learning from any a priori fixed class distribution. ACS is the process of iteratively selecting class proportions for data generation.

- **Training set:** Use the same training set in Phase I.
- **Test set:** Simpson, Karen Elizabeth's group does the experiment on tile1_24 and tile1_25. Stillman, Christopher A's group does the experiment on tile2_24 and tile2_25. Maio, Christopher V's group does the experiment on tile3_25.
- Use four methods listed below to do Active Class Selection; read Paper "Active Class Selection" Section 2 ACS Methods for detailed explanation. As to the first three methods, choose several instances from the training set as the original training set, then use J48 classifier and perform 10-folder cross validation on the training set to get the prediction result. Use the equation of every method to add instances to the original training set n times. In these equations, b[r] is the number of instances you add in round r (r<=n). Pr[c] is the class proportion in class c. we have two classes: crater and non-crater. The number of instance in original training set and b[r] can be chosen randomly or any number you want. Perform n times until the whole training set is used. As to the 4th method, do the second round the same as method 1 or 2 or 3, then perform the following round using the equation in 4th method.

1. First method : Uniform

   $\Pr[c] := \frac{1}{2} * b[r]$

   In this method, add ½ *b[r] to each class in round r

2. Second method: Inverse

$$\Pr[c] := \frac{\frac{1}{acc\,[c]}}{\sum_{i=1}^{|classes\,|}\frac{1}{acc\,[i]}} * b[r]$$

acc[c] is the accuracy of class c.

In this method, select class proportions *Pr*[*c*] inversely proportional to their CV accuracy on round *r* – 1. Thus, we obtain more instances from classes on which we have a low accuracy. This method relies on the assumption that poor class accuracy is due to not having observed sufficient training data.

Add $\frac{\frac{1}{acc\,[c]}}{\sum_{i=1}^{|classes\,|}\frac{1}{acc\,[i]}}$ *b[r] to each class in round r.

3. Third method: Original Proportion

   $\Pr[c] := n_c * b[r]$

   Sample in proportion to the class proportions in training set *Tr*. The idea is that domain knowledge led to these proportions, perhaps because they are the true underlying class distribution or because of the creator's intuition as to which classes are more difficult. $\Pr[c] := n_c * b[r]$, where n_c is the proportion of class *c* found in the collected data *Tr*.

   Add $Pr[c] := n_c$ *b[r] to each class in round r.

4. Fourth method: Accuracy Improvement

   $$\Pr[c] := \max(0, \frac{currAcc[c] - lastAcc[c]}{\sum_{i=1}^{|classes\,|} currAcc[i] - lastAcc[i]} * b[r])$$

   currAcc[c] is the accuracy of current training set, lastAcc[c] is the accuracy of last training set.

   Sample in proportion to each classes' change in accuracy from the last round. If the change for class *c* is ≤ 0, then *Pr*[*c*] = 0. The intuition is the accuracy of classes that have been learned as well as possible will not change with the addition of new data and thus we should focus on classes that *can* be improved. This method looks for stability in the empirical error of each class.

   Add $\max(0, \frac{currAcc\,[c] - lastAcc\,[c]}{\sum_{i=1}^{|classes\,|} currAcc\,[i] - lastAcc\,[i]} * b[r]$ to each class in round r.

- Use these methods in Weka J48 and draw a graphic as the Fig1 in the "Active Class Selection" paper. The x value is total number of instance you used as the training set and y value is the accuracy of every round on the test set.

Submission Requirements

1. 3 teams use 3 different data set for testing. 3 teams may observe different performance of the 4 ACS methods discussed above. Write an experiment report to discuss all the experiments. In the report, discuss the design of experiments, the parameter setting, experiment results and your conclusion. Evaluate your experimental results carefully and discuss any interesting results.

2. 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.
3. Submit the softcopy of the report and scripts through your UMassOnline account at http://boston.umassonline.net/index.cfm.
4. 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.*
5. No hard copies or soft copies results in 0 points.