CSCI 578: Machine Learning - CSCI 481: Data Mining
MARS Mineral Discovery Challenge

Background:

The identification of phyllosilicates by NASA's CRISM (Compact Reconnaissance Imaging Spectrometer for Mars) strongly suggests the presence of water-related geological processes. A variety of water-bearing phyllosilicate minerals have already been identified by several research groups utilizing spectral enrichment techniques and matching phyllosilicate-rich regions on the Martian surface to known spectra of minerals found on earth. However, fully automated analysis of the CRISM data remains a challenge for the following reasons.

1. There is significant variability in the spectral signature of the same mineral obtained from different regions on the Martian surface.
2. Some minerals appear with low abundance making collection of large number of training regions from these minerals challenging, thus leading to a highly imbalanced training dataset.
3. The list of mineral confirmed to date constituting the set of training classes is not exhaustive. Thus, when classifying new regions, using a classifier trained with selected minerals one must consider the potential presence of unknown materials not represented in the training library.
4. CRISM images contain lots of ambiguous and uninteresting pixels that should not be classified because these pixels cannot be identified with certainty. If these pixels are incorrectly classified into known minerals they obfuscate otherwise detectable spectra of these minerals. Any classifier should also deal with the problem of when to classify a pixel and when not to classify it and call it an outlier.

Challenge:

Challenge participants are asked to build machine learning algorithms using the released training data. The trained classifiers will be evaluated based on their predictive performance in classifying pixels in the test data set.

Datasets and Evaluation Criteria:

Several hundred CRISM images are surveyed over the past few years to construct an extensive training library of around 30 mineral classes. Each mineral has at least one occurrence but some has as many as dozens. The training library contains the following variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Dimension</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>pixspec</td>
<td></td>
<td>Spectra for each training pixel</td>
</tr>
<tr>
<td>pixims</td>
<td></td>
<td>Vector of image IDs from which the spectrum is extracted.</td>
</tr>
<tr>
<td>pixlabs</td>
<td></td>
<td>Vector of numeric mineral labels assigned to each pixel.</td>
</tr>
<tr>
<td>pixcrds</td>
<td></td>
<td>Coordinates of the pixels in the image</td>
</tr>
<tr>
<td>pixpats</td>
<td></td>
<td>Vector region IDs</td>
</tr>
</tbody>
</table>
Mean F1 score will be used to evaluate the performance of the predictive models. Training data will be released early March. The test data will not be released until after Spring break. Submission window will be open for the week of April 6 and competition will last for four weeks. Each participant can have three submissions per week.

**Directions for Literature Review:**

This data set is interesting on multiple fronts. CS 481 students are expected to tackle the first two challenges. CS 578 students are expected to tackle at least three of the four challenges.

- How to model inter- and intra-class variability?
- How to deal with class imbalance and rare classes?
- How can we incorporate classes with lab spectra but no detected real-world spectra into training?
- How can we implement an open-set classifier and avoid classifying uninteresting pixels?

**Timeline:**

First week of March: Training data set released

March 20: Proposal are due

March 27: Test data is released

April 6: Competition begins

May 1: Competition ends.