Classification and Filtering of Spectra: A Case Study in Mineralogy

Jonathan Moody
Computer Science Department
Carnegie Mellon University
jwmoody+@cs.cmu.edu

Ricardo Silva
Center for Automated Learning and Discovery
Carnegie Mellon University
rbas+@cs.cmu.edu

Joseph Vanderwaart
Computer Science Department
Carnegie Mellon University
joev+@cs.cmu.edu

Joseph Ramsey
Philosophy Department
Carnegie Mellon University
jdramsey+@andrew.cmu.edu

Clark Glymour
Philosophy Department
Carnegie Mellon University, and
Inst. for Human and Machine Cognition
University of West Florida
cg09+@andrew.cmu.edu

December 16, 2002

Abstract

The ability to identify the mineral composition of rocks and soils is an important tool for the exploration of geological sites. Even though expert knowledge is commonly used for this task, it is desirable to create automated systems with similar or better performance. For instance, NASA intends to design robots that are sufficiently autonomous to perform this task on planetary missions. Spectrometer readings provide one important source of data for identifying sites with minerals of interest. Reflectance spectrometers measure intensities of light reflected from surfaces over a range of wavelengths. Spectral intensity patterns may in some cases be sufficiently distinctive for proper identification of minerals or classes of minerals. For some mineral classes, carbonates for example, specific short spectral intervals are known to carry a distinctive signature. Finding similar distinctive spectral ranges for other mineral classes is not an easy problem. We propose and evaluate data-driven techniques in two stages: first, evaluating algorithms to identify which components are probably present in a given rock; second, trying to improve this classification by automatically searching for spectral ranges optimized for specific classes of minerals. In one set of studies, we partition the whole interval of wavelengths available in our data into sub-intervals, or bins, and use a genetic algorithm to evaluate a candidate selection of subintervals. As an alternative to these computationally expensive search techniques, we present an entropy-based heuristic that gives higher scores for wavelengths more likely to distinguish between classes. Results are presented for four different classes, showing reasonable improvements in identifying some, but not all, of the mineral classes tested.

keywords: scientific applications, classification, filtering.

1 Introduction

Reflectance spectrometers have been used for identification of mineral composition of rocks and samples with varying degrees of success. This kind of spectrometer measures the amount of sunlight reflected by a rock or soil sample over a range of wavelengths. The reflectance pattern obtained under different wavelengths can then be used to predict which minerals are present in that sample.

For instance, NASA intends to design robots for planetary exploration that would be sufficiently autonomous to interpret spectrometer data and report only the results back to Earth. Robots equipped with automatic classifiers of rock and soil samples would also be useful for automatically planning which different regions of a geological site would be more promising for prospecting certain classes of minerals.

So far as planetary exploration is concerned, reflectance spectroscopy techniques have already shown themselves to be useful. Visual to near infrared (VNIR) reflectance spectroscopy (from approximately 0.4 μm to approximately 2.5 μm) in particular has offered geologists an important potential source of petrological information for the exploration of planets, satellites, and other solar system objects. Lightweight, low-power commercial instrumentation is available, detailed physical models have been developed (e.g. [9]), and data from VNIR instruments is routinely used by geological spectroscopists in practical mineral classification.[1] Were such instruments coupled with intelligent software for mineral classification from spectra, the resulting system could be used either for remote sensing or for surface based studies, reducing requirements for data storage and information transmission, and aiding autonomous, rational, scientifically-informed decisions by

[1] See, for example, Chapters 3, 14, 16, 20, and 21 of [13] and references therein.

robot explorers about further directions for exploration and data acquisition.

As one specific application, this interest in planetary exploration motivates an examination of the problem of determining whether rock or soil samples contain carbonates and, in particular, whether such samples contain either of the most frequently occurring forms of carbonate material-calcite or dolomite. Carbonate identification is interesting for extra-terrestrial exploration, because carbonates are typically formed by processes, such as deposition from water, which could indicate a history of an environment that once supported life.

The data sets collected by spectrometers consist of levels of reflectance intensity of a given rock at different wavelengths. In Figure 1 we have a plot of the intensity of reflectance of a particular mineral at different wavelenghts. The intensity data are typically measured relative to a reference surface in order to be invariant with respect to the total amount of sunlight in the environment.

The usual approach taken by someone interested in building a predictive model out of this data is running a regression model for each rock or soil sample, where the dependent variable is the reflectance intensity of the unknown rock and the independent variables are the reflectance intensities of a variety of different pure minerals that are possible components of the rock, measured over the same wavelengths. Libraries of such pure mineral spectra exist; in particular, the Jet Propulsion Laboratory (JPL) has produced a library of spectra for 135 different pure minerals, each containing reflectance intensities for 820 different wavelengths between 0.4 and 2.5 μm. We will refer to the JPL library in experiments described later in this paper.

Assuming that the intensity of the rock is a linear combination of the intensity of its components, a regression model is built using each reflectance value at a wavelength as a data point. Then, only those minerals whose coefficients on the regression model pass a given test of statistical significance are considered components of the rock. A successful learning algorithm should commit as few errors as possible, where an error is accepting a given mineral as part of a rock when this is not true, and rejecting a given mineral as part of a rock when in fact it is. In the following sections we will introduce a modified regression approach that gives better results than the standard regression and many other machine learning algorithms, compare it with some available results of human experts and expert systems, and introduce a second level of automation by developing search algorithms for intelligent data filtering.

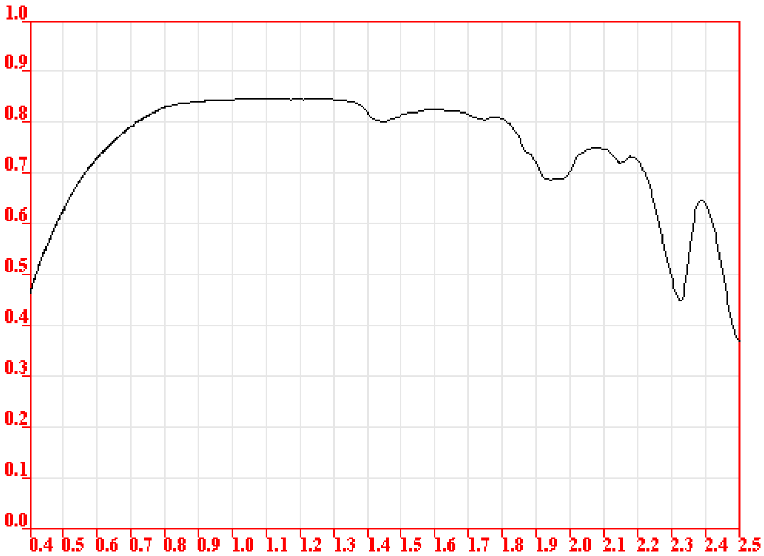

Figure 1: The reflectance spectrum profile of a specific mineral from the class of carbonates. The horizontal axis corresponds to different wavelenghts ranging from 0.4 to 2.5 μm. The vertical axis is a scaled measured of reflectance intensity.

2 A new method for identification of mineral components

The identification of surface composition of rocks from reflectance spectra has traditionally relied on two methods. The older of the two is a direct examination of spectra by experts, seeking lines or bands characteristic of particular substances, sometimes taking account of overall luminosity of the spectrum, and sometimes, with computational aids, taking account of the shapes of bands. The standard alternative is simultaneous linear regression of an unknown spectrum against a library of known spectra for candidate materials. A number of spectral libraries have been compiled which can be used for this purpose. Some neural net procedures—notably Kohonen maps—have also been used to analyze spectral data, typically not for identifying surface composition directly, but rather for finding bounded regions of similar composition in an array of point spectra from a "visual" field.[2] Other automated techniques have been used explicitly to identify surface composition of minerals and rocks, including a Bayesian technique described below. Despite, however, its numerous applications for planetary and terrestrial exploration and for various military purposes, we have found no published systematic (or even unsystematic) study comparing automated examination of reflectance spectra to human expert examination of reflectance spectra.

[2] Careful work on this subject has been carried out by E. Merenyi—e.g., [12].

In [16], we compared the reliabilities of (a) an expert human spectroscopist,(b) an expert system that models human expert procedures, and (c) a variety of automated techniques, including linear regression, each with various resampling and cross-validation techniques, on the task of carbonate identification from visual to near infrared reflectance spectra. All of our tests of data mining procedures use the same library of spectra for training or reference. A variety of data sets are used for testing, including laboratory and field spectra obtained under various conditions. The following subsections summarize part of the results obtained.

2.1 Component detection approaches

When applying regression to this problem, each independent variable is the spectrum measures of each mineral from the reference library (such as the JPL library), and the outcome variable is the spectrum measure of the rock that should be classified. Under the assumption of normal additive error, the coefficients in the estimated regression model have a well-defined probabilistic distribution. Each coefficient is evaluated by a hypothesis test, and only those that are statistically significant are considered.

The minerals in the library are grouped in pre-determined classes, such as carbonates and oxides. The grouping of minerals is a domain dependent classification. We classify a rock as containing elements of class **T** if at least one of the minerals of class **T** in the reference library has a significant coefficient in the estimated model. Notice that, unlike typical predictive approaches for classification, we estimate a new model for each new case (rock) that shoud be classified.

The standard regression suffers from some difficulties when applied to to inverse problems such as identification of components in a mixture of signals:

- if there is a common cause between two regressors X and Y that is not included as a third regressor, and one of the regressors is significant but the other is not, both will be significant if the sample size is large enough. This will bias the regression model to include components that should not be considered. The phenomenon is sometimes called conditional correlated error. In the present application, it can result in the identification of minerals that are not, in fact, components of the source.

- simultaneous linear regression computes the partial regression coefficient of a variable X by conditioning (assuming a Normal distribution) on all other regressors. In our application, conditioning on all of the other minerals in the reference library. While any one of these variables may be only loosely correlated with X, together they may be highly correlated with it. In that case, the significance of X may be effectively zero. In the present application, multicollinearity can result in failing to identify a true component of the source.

- the variance of the estimates of a simple regression coefficient is a function of the sample size and the number of other candidate causes, or regressors. The bigger the sample size and the smaller the number of other regressors, the smaller the variance. Assuming a Normal distribution, the trade-off is one for one: adding an extra regressor variable is equivalent in its effect on the variance to reducing the sample size by one unit. In the present application, reducing the number of channels used for data analysis increases the variance of the estimates of regression coefficients. In the extreme case in which the number of variables is greater than the sample size, regression is ill-defined, and standard regression packages will not run at all.

We could use a stepwise regression procedure, but other experiments with small samples have found stepwise regression less reliable than the PC algorithm used in [21], which will then be adapted to attend the special necessities of our application.

All three of the problems cited above stem from a single structural feature of the regression procedure, linear or otherwise. Let $\mathbf{C}$ be our of regressors. In estimating the influence of a variable X on the outcome Y, regression conditions simultaneously on all other candidate variables, i.e., all of the other members of $\mathbf{C}$. That is, in our (rather conventional, but not textbook) use of regression, we test the null hypothesis that X has no influence on Y (or is not a component of Y) by using the distribution of a test statistic that is conditioned on all other members of $\mathbf{C}$.

There is an alternative procedure that minimizes the number of variables that must be conditioned on. It takes as input a set of background variables $\mathbf{C} = \{X_1, X_2, ..., X_n\}$ together with a target variable Y not in $\mathbf{C}$ and dynamically eliminates variables from $\mathbf{C}$ using conditional independence facts, calculated from data. Variables are eliminated which are independent of Y conditional on subsets of other remaining variables in $\mathbf{C}$, where the cardinality m of the subsets increases in size (m = 0, 1, 2, ...) until no more variables can be eliminated from $\mathbf{C}$. More formally:

Modified PC Algorithm: Given set C of background variables and target variable Y:

1. for each X_i in $\mathbf{C}$, test the hypothesis that the correlation of X_i with Y is zero; if the correlation of X_i with Y is zero, $\mathbf{C} := \mathbf{C} - \{X_i\}$;

2. for each X_i in $\mathbf{C}$, and for each $X_j \neq X_i$ in $\mathbf{C}$, test the hypothesis that the correlation X_i with Y, controlling for X_j , is zero; if this is true let $\mathbf{C} := \mathbf{C} - \{X_i\}$;

3. for each X_i in $\mathbf{C}$ and each X_j, $X_k \neq X_i$ in $\mathbf{C}$ test the hypothesis that the correlation X_i with Y, controlling for $\{X_j, X_k\}$ is zero; if this is true, let $\mathbf{C} := \mathbf{C} - \{X_i\}$;
   ...and so on, until no more members of $\mathbf{C}$ can be removed. Return $\mathbf{C}$.

If n members of $\mathbf{C}$ are actually components of Y, no more than n variables must be conditioned on simultaneously. If, for example, three minerals in the JPL library are actual components of a sample, a large number of statistical tests will be done, but none of the tests will require controlling for more than three variables. In no test will the sample size effectively be reduced by more than 4, in contrast to multiple regression in which the sample size is reduced by 134 (since there are 135 minerals in this library, what accounts for 135 regressors). For that reason, unlike multiple regression, the procedure can be used with the JPL library with the reduced data set using only intensities in channels for wavelengths in the interval [2.0 μm, 2.5 μm]. The importance of reducing the range of used wavelengths will be discussed later in more detail.

2.2 Evaluation

To test the performance of the modified PC algorithm, we need a library of reference spectra for minerals (in our case, we used the JPL library) and a data set of measurements of rocks which composition will be classified. The Johns Hopkins University (JHU) has assembled a library of reflectance spectra for a variety of solid and powdered rock samples. Each spectrum in the JHU rock library is accompanied by a description of the petrology of the sample. Because mineralogical nomenclature is so varied, these descriptions do not generally identify sample components either as among the 135 specific minerals represented in the JPL library (e.g., calcite, dolomite, etc.) or as among the 17 general mineral classes into which the JPL library is classified (e.g., carbonates, phyllosilicates, etc.). Assignment of JHU samples to the 17 general JPL mineral classes on the basis of the petrological descriptions alone requires expert knowledge.

Using the JHU petrology descriptions, but without access to the sample spectra, Ted Roush of NASA Ames determined which of the 17 JPL mineral classes is represented in each of the 192 JHU rock samples. Since the rocks were not pure minerals, they could each belong to more than one of the 17 general mineral classes. Ramsey [16] performed a exhaustive detailed comparison of the modified PC procedure, an expert system, a human expert and the tools contained in the Model

1 software. Model 1 includes linear regression, logistic regression, feedforward neural networks, CART, naïve Bayes, and other procedures.

The main task was identifiying if a rock contained carbonate minerals. According to Roush's classification, 92 of the samples were judged to contain some form of carbonate. These assignments of JHU minerals to carbonate class were then used as ground truth in tests of reliabilities of various procedures for mineral classification. When applied to raw measurements, none of the procedures was much better than a random choice, but the Modified PC was still clearly better than regression and the best model chosen by Model 1.

In order to build actually useful models, some kind of preprocessing of the data was necessary. One step was smoothing the measurements. Other particularly helpful step was using some expert knowledge to select subparts of the spectrum range that would be more informative. With this preprocessing, the Modified PC algorithm was able to reach approximately 63% classification accuracy against a baseline of roughly 50%. Detailed account of these choices are given in [16] and [17].

3 Another level of improvement: intelligent data filtering

Experiments with the specific class of carbonates have shown that restricting the input of the PC algorithm to a smaller region of the spectrum can improve accuracy. In particular, a region suggested by prior expert knowledge (a region used by experts to identify carbonates) produces much better results than allowing the algorithm to consider the entire spectrum. In other words, the filtered spectrum does not include noninformative or noisy data that could confound mineral identification. This is a promising result that arguably can be extended to other classes.

Carbonates show a very typical curve on the spectra region between 2.0 and 2.5 μm, which motivated the scientists to focus on this region. However, coming up with a good range of wavelengths is not an easy task because little is known for other mineral classes. No automated method has been applied in [16] to find subintervals that would be more appropriate for identifying given classes and subclasses of minerals.

Our goal was to find intervals of the spectrum, specific to each class of minerals, for which the PC algorithm performs better than the same algorithm using the entire spectrum. This is a search problem that complements other data pre-processing issues described in Section 5. We tried several methods, a collection representative of both heuristic and computational intensive approaches that also bear relation with feature selection techniques.

4 Data filtering techniques

Finding an appropriate subset of the spectrum range can be cast as a problem of search among the space of possible subsets. Since we have over 800 available channels, an exhaustive search is infeasible. Also, a larger number of evaluated candidates increases the chance of overfitting [3]. One must decide how to trade-off the complexity of the search space depending on the chosen search algorithm, the available computational resources, and the amount of data available.

By the terminology used in feature selection research, as described in [10], we are basically building wrappers over the PC algorithm. Four algorithms were tried: a computationally demanding genetic algorithm, two greedy hill-climbing algorithms and a simple grid search strategy over a rather reduced number of parameters of a customized evaluation function.

The data filtering methodologies described here should be applied to each class of minerals at a time, since a interval that is suitable to one class is unlikely to be useful to other.

4.1 Genetic algorithm

A genetic algorithm is an algorithm for combinatorial optimization [6], which is directly related to the task of finding useful subsets of the spectra. The most straightforward representation of a candidate is through a string of 826 bits, where a positive bit represents that the respective channel will be used. However, due to the reasons explained in the beginning of this section, we divided the spectrum into a fixed number of blocks, each represented by a bit. Thus, all channels in the same block are selected or not selected at the same time.

The evaluation function is very time-consuming: it consists in running the modified PC algorithm over a whole set of rock samples. The fitness of a candidate is the proportion of rocks that are correctly classified as containing or not containing the respective mineral. On our available implementation, it takes about 30 seconds to evaluate a single candidate feature mask on a Pentium III 733MHz processor.

4.2 Bitwise hill-climbing

We also used a greedy, hill-climbing algorithm that uses the same representation for search states and the same evaluation function. On the initial state, all bits are activated. The next states are generated from the current state by setting to zero one of the currently activated bits. If the current candidate has n activated bits, it will generate n new candidates. The candidate with the highest evaluation value is chosen to be the next state.

4.3 "Peeling" algorithm

This is another greedy algorithm that is also used for rule induction over continuous/ordered attributes [5]. It consists of trimming the extremes of an interval by some percentage of the data and evaluating the new interval obtained. A typical strategy starts with the complete interval and, at each subsequent step, generates three new candidates: the current interval with the bottom $\alpha\%$ of the ordered data discarded, the current interval with the upper $\alpha\%$ of the ordered data discarded, and an interval constructed by dropping the bottom and upper $\frac{\alpha}{2}\%$ from the current interval.

The underlying assumption of this algorithm is that interesting intervals are continuous. Unlike the previous algorithms, all selected subintervals are of the form $[a, b]$, where a and b are points of the original interval. It may clearly result in suboptimal selections, at the advantage of being much less time demanding.

4.4 Information gain heuristic

A more straightforward approach would be to construct a "relevance" heuristic, rank the channels accordingly, and select those with relevance above a threshold. Intuitively, we wish to discover those channels that carry a large amount of information relevant to the question of whether a certain class of minerals is present. Therefore, we used information gain, a quantity based on entropy, for our relevance heuristic.

The information gain algorithm for selecting a channel mask is as follows. For each channel, we divide the intensity range into some number of bins. Then for every spectrum in the reference library we look at the intensity at the current channel and take note of which bin it occupies and whether or not it is a member of the target class. When we have finished doing this for a given channel, we calculate the fraction of samples in each bin that are in the desired class; this number is used to calculate an entropy value for that bin. A weighted sum of the entropies of the bins gives the expected entropy given the intensity of a particular channel; subtracting this from a constant gives the *expected information gain* associated with that channel.

When we have calculated the expected gain for each channel, we create a channel mask by looking for intervals where the expected gain is higher than average. Specifically, we divide the

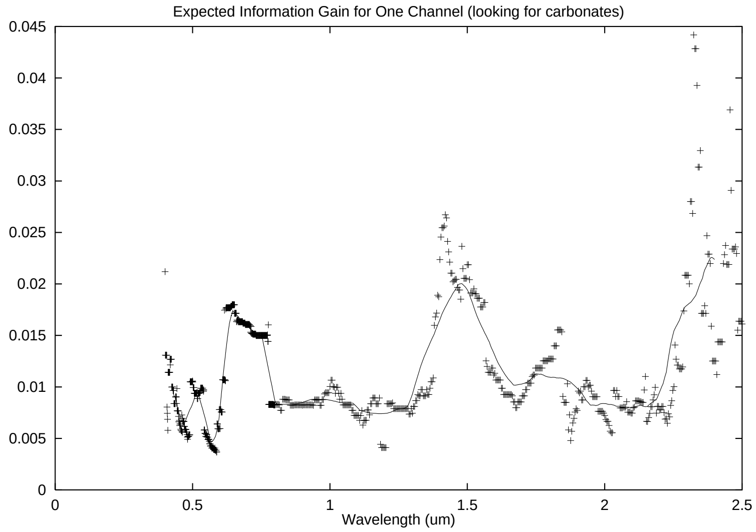

Figure 2: The average information gain per each channel using four bins with respect to the carbonate class. Notice that the highest gains lie on the upper region of the spectra, as suggested by expert knowledge.

spectrum into blocks and calculate the average expected gain in each block. Then blocks whose average expected gain exceeds the global average by some margin are selected for use in classification.

Under this technique, we optimize the number of bins and threshold parameters by performing a grid search over a given interval of possible values. The selection that gives the best classification accuracy for the training set is used. Figure 2 shows how it is possible to visualize promising regions using this evaluation function.

5 Experiments

For our experiments we used the NASA Jet Propulsion Laboratory (JPL) data set as a reference library, attempting to classify the rocks in the Johns Hopkins University (JHU) data set [3] Each mineral on JPL was measured with different grain sizes. We used the largest grain size, which should give a closer approximation to rocks found on test fields. The data set was processed to subtract the effect of environment luminosity by considering measures of reflectance relative to a white surface. It was necessary to interpolate the measures of the JHU spectra in order to match

[3]Documentation for these data sets can be found in `http://speclib.jpl.nasa.gov/`

the same wavelengths found on the JPL library.

Also, most features of spectra which are diagnostic of the chemical structure of minerals are small scale "dips," or deviations, from the overall background shape of the spectrum, with a width on the order of 1 to 50 μm. By taking the hull difference of a spectrum, variations due to the large-scale shape of the spectrum are reduced or eliminated and variations due to these smaller, typically more diagnostic, variations are enhanced. On the following experiments, we refer to data treated by the hull difference process as the "processed data", while "raw data" will refer to spectra without this modification. For further information on these data sets, see Ramsey *et al.* [16].

We performed experiments using four of the mineral classes available in the JPL library. These minerals were chosen according to the number of rocks present in the JHU data set that were reported to have these minerals: it would be unreliable to try to find intervals for a class underrepresented in the available data. Among all 192 JHU rocks, 92 have carbonates, 121 have phyllosilicates, 100 have oxides and 84 have inosilicates.

Tables 1 and 2 show the results for running the modified PC algorithm using the intervals selected by variations of each algorithm described on the previous sections. For each mineral class, we ran a five-fold cross-validation. The accuracy measure is the number of correctly classified rocks (true positives plus true negatives) divided by the number of rocks on the corresponding sample.We opted for 5 folders instead of the usual 10 folders because:

- the genetic algorithm is computationally intensive;
- we wanted a reasonable amount of data on both training and test sets. Using a high number of folders can in fact lead to worse generalization estimates when we have few data points and the prediction error is high, as it is typical of this domain [20].

The whole spectrum interval was divided in 15 subintervals of equal size. For the genetic algorithm, for example, this means we are using 15 genes per individual. The reason for this choice was to allow approximately 50 wavelengths per cell and to avoid introducing too much variation on the search for selected intervals. A more extensive experimental analysis could include this choice as a parameter to be optimized.

For the genetic algorithm, we used 35 individuals. The search proceeded for at most 40 generations. In all cases, by the last generation the pool of individuals was almost completely dominated by copies of a single individual (and in many cases, all individuals were identical), suggesting that

	None	GA	HC	PEEL	IG
Carbonates	56.3 ± 8.6	64.0 ± 5.0	62.0 ± 4.2	52.7 ± 13.0	66.2 ± 6.9
Inosilicates	61.4 ± 8.3	69.7 ± 6.7	65.5 ± 7.5	60.0 ± 7.5	70.0 ± 8.1
Oxides	56.3 ± 6.7	58.7 ± 6.7	49.9 ± 3.1	48.9 ± 7.5	56.3 ± 4.0
Phyllosilicates	56.3 ± 7.5	57.1 ± 3.8	50.2 ± 5.6	55.7 ± 2.1	50.0 ± 5.2

Table 1: Mean and standard deviation for classification accuracy (in %) obtained by using the raw data. GA stands for genetic algorithms, HC for the general hill-climbing algorithm, PEEL for the "peeling" procedure and IG is the label for the information gain results. The first column represents the results obtained when all the spectrum is used.

	None	GA	HC	PEEL	IG
Carbonates	63.4 ± 7.0	68.3 ± 3.9	66.1 ± 5.3	65.5 ± 5.2	61.4 ± 7.6
Inosilicates	61.4 ± 4.3	66.3 ± 4.3	68.4 ± 4.0	66.1 ± 11.1	60.0 ± 10.9
Oxides	49.3 ± 6.1	48.5 ± 1.4	53.7 ± 9.6	50.0 ± 5.1	50.9 ± 5.9
Phyllosilicates	54.1 ± 6.0	52.7 ± 6.1	53.7 ± 7.5	53.7 ± 6.8	59.4 ± 3.1

Table 2: The results obtained for the processed data.

further optimization would not improve the result obtained significantly. The code of the genetic algorithm was adapted from [11], with its default parameters.

We also used cached statistics to scale up the algorithm: instead of passing through all the data points when computing an element of the correlation matrix (as required by the PC algorithm), we precomputed the summations and inner products of variables for the data falling under each block. Getting a new element of the correlation matrix required only a pass over these cached statistics. This procedure reduced the computational time by over 30%.

For the standard hill-climbing search, we adopted the following stopping criterion: as a trade-off to avoid bad local maxima without searching till the last state, the search stopped when we did not get improved results for five consecutive states. The best selection on this search path was the output.

For the peeling algorithm, we used a value of 5% for α. We used the same stop criterion applied on the previously described hill-climbing technique.

To find appropriate parameter values for the entropy heuristic, each training set was used to evaluate the masks produced by several different parameter settings. In particular, all possible combinations of 3, 4 or 5 bins with thresholds of 0.1, 0.2, 0.3, 0.4 or 0.5 standard deviations above the mean gain were tried. For each training set, the mask that produced the best accuracy was selected as the optimal mask and its fitness was measured with the corresponding test set.

Using the interval selected by experts, we get an accuracy of 67.7% for the raw data and 66.1% for the processed data. By comparison with the results obtained, it is clear that some of our approaches were overall able to find selections with similar performance, but unable to significantly improve over it. We should not forget, however, that these results were attained without relying on background knowledge and hence provide evidence that for cases where this knowledge is actually unavailable, this set of approaches can be a useful tool.

The data pre-processing by taking hull differences can help in some occasions, as was the case for carbonates. For the inosilicates, however, reasonably better results were obtained using the raw data. As with any smoothing procedure, it can be useful in some situations, but not always. The information gain heuristic proved especially sensitive to this technique.

While our performance on carbonates and inosilicates improved relative to the baseline of enabling all channels, we got unimpressive results with phyllosilicates and oxides. It was expected that for some classes the reflectance spectrum information is not sufficient to provide a good separation between those classes and the remaining ones. In this ill-defined situation, data filtering

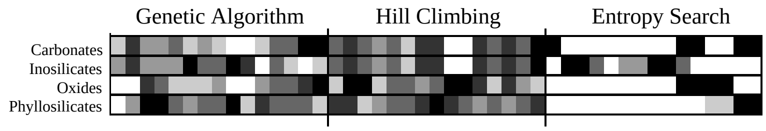

Figure 3: This figure depicts the number of times each of the fifteen cells was chosen across the five training sets used in the cross-validated experiments. A cell that is totally black was chosen all times, while a white space represents a cell that was never chosen. It is interesting to notice that the entropy search was the most stable, but that the exact selection of a given mask is not required for reasonably good generalization.

would not be able to help much.

The variance of the results is due not only to sample variance, but also to the variance of the underlying classifier, the simplified PC algorithm. Depending on the data selection algorithm, we have also small or big variance on the selected intervals. Figure 3 depicts the number of times each cell was chosen for some of the algorithms on the raw data. Due to its simplicity and reduced number of parameters, the entropy heuristic was the most stable.

Spectral libraries for rocks and minerals, respectively, are also available for portions of the infrared range. The Johns Hopkins University Spectral Library contains a set of 160 minerals grouped into 12 major mineral categories in the range 2.1 - 25 μm. The library also contains the collection of rock samples used above in the range 0.4 - 2.5 μm, although the spectra themselves extend out to 14.0 μm. Those rock samples were classified by an expert geologist (T. Roush) into the same 17 mineral categories used by the Jet Propulsion Laboratory Spectral Library; many of these same categories appear as categories for the 160 minerals of the JHU Spectral Library. So in the range 2.1 - 14.0 there are mineral and rock spectra that at least with some initial plausibility could be used to automatically generate domain masks.

Examining rock and mineral spectra even in this restricted infrared range (2.1 - 14.0 μm) would be of some benefit. Salisbury *et al.* ([19], p. xiv-xv) discuss wavelength bands and ranges useful for identifying a number of different mineral classes. For instance, the wavelength range 8 - 12 μm contains strong spectral features that are characteristic of silicates, carbonates have bands near 7 μm and μm, and sulfates have a distinctive band at 8.7 μm. It is likely that an automatic mask-construction algorithm would verify these expert geological observations and suggest other wavelength ranges of interest as well, for these and other mineral classes.

	None	GA	HC	IG
Carbonates	72.3 ± 4.3	76.5 ± 3.9	77.9 ± 9.0	76.0 ± 6.2
Inosilicates	59.0 ± 4.7	69.2 ± 5.9	65.5 ± 6.5	68.7 ± 7.8

Table 3: Mean and standard deviation for classification accuracy (in %) obtained by using data that includes infrared range up to 14 μm. Again, the first column represents the results obtained when all the spectrum is used.

We performed extra experiments with a JHU library of minerals that includes measurements in the infrared range for the carbonate and inosilicate cases. Table 3 compares the results of the three main algorithms discussed here. Due to the higher number of available wavelengths, we decided to split the interval in 25 bins instead of the 15 used in the experiment discussed above.

Reasonable improvements could be also detected for these two classes, especially for inosilicates. It is also important to point that simple algorithms such as the entropy heuristic were competitive when compared with the genetic algorithm. Since our data sets were small, computational time was not a major issue, but in applications where a larger number of measurements is performed, they can turn out to be viable solutions.

6 Related work

The techniques applied in this work are related to the areas of feature selection and data cleaning. Wettschereck, Aha and Mohri [23] formulate a framework for feature weighting methods under the context of lazy learning. Even though in a strict sense the wavelength channels are in fact rows of our data set, not attributes, in principle one can use these techniques to weight the relevance of each data point (or intervals for practical purposes). According to the categories of Wettschereck *et al*'s framework, the genetic algorithm and hill-climbing approaches would be classified as having:

- a performance bias, since we use the actual results of classification for deciding the selection;
- a binary weight space (*i.e.*, 0/1 weights);
- a transformed representations, since we divide the data into blocks;
- a global weighting, because the same intervals are selected for all minerals;

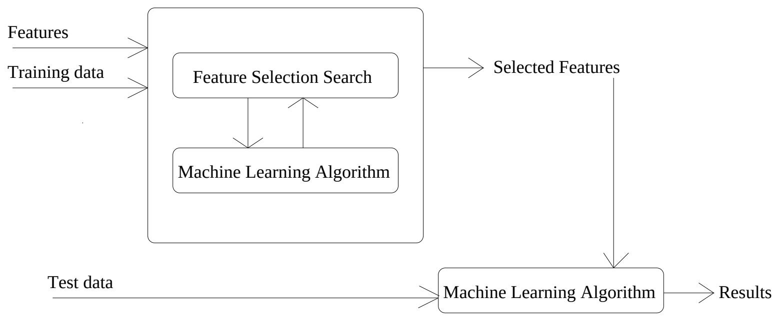

Figure 4: the "wrapper" approach

- knowledge-poor, since we did not use prior knowledge in our experiments. Hull differences help on some situations;

The performance bias is also commonly described as a wrapper approach [10]: our selection policies use the modified PC algorithm as a black box that outputs a measure of performance (Figure 4).

Unlike general feature selection problems, we do not have the concern of selecting features that present fewer missing values on the available data bases, nor do we have to consider which are more expensive to measure (e.g., some medical exams for diagnosis problems). That makes our fitness function even simpler than most ones used in feature selection literature [14, 22, 24]. These approaches are virtually identical to the genetic algorithm for data selection described in this work, where the difference is mainly a more complicated evaluation function. Demiroz and Guvenir [4] also describe mechanisms for learning continuous weights between 0 and 1, which arguably are not very useful for our problem, where we have too little data to accomodate such a precise tuning of parameters.

In contrast, the information gain heuristic operates as hybrid between a wrapper and a filter approach (Figure 5). The filter approach applies for each feature a measure of importance that is independent of the learning algorithm that will be used. Hall [8] provides a comparison of filters and wrappers, as well as an overview of feature selection. He favors the filter approach due to its much higher scalability, but in his discussion it is mentioned that ideally the features themselves should be a function of the bias of the learning algorithm that will be used. An intermediate approach such as using the entropy measurements to search for a combination of prominent intervals, which can then be successfully used by the modified PC algorithm, is a way to trade-off these issues.

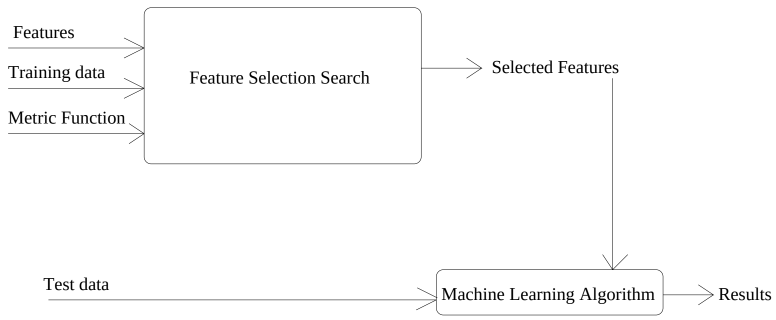

Figure 5: the “filter” approach

Entropy measures are commonly related to the degree of unexpectedness of a pattern, and such a characteristic has been explored for data set cleaning. Guyon, Matic and Vapnik [7] describe different ways of using information theoretical measures to identify outliers or highly informative examples. Data points are ranked according to information gain and then submitted to a expert that will classify them as outliers or representative examples. Guyon *et al.* warn against the risk of getting improved results during training by dropping the most difficult examples and then achieving bad generalization accuracy.

Another application of information theoretical measures for data cleaning is discussed by Pyle [15], where it is also described how to find ill-defined regions of a function by checking symmetries between the input and output variables. This specially affects inverse function estimators. Pyle also describes what he calls “attention processing” of data: how to perform data surveying and avoiding on the combinatorial explosion of the search space of potentially problematic regions of the data.

7 Conclusions and future work

The experiments provided some evidence that approaches such as genetic algorithms can improve the classification performance for this domain. However, sampling variability may be a concern and the fact that the underlying classifier provides its own source of variability may amplify this problem. Kohavi and George [10] report that feature selection algorithms may overfit easily. Approaches to minimize this problem and perform more reliable performance assessment include resampling techniques such as bootstrapping [2].

For example, it may be possible that more robust masks of selected intervals can be obtained by the combination of different masks. One simple policy is obtaining multiple masks by resampling and then giving to each bin a weight proportional to the number of times each one appears. That changes our problem from "feature" selection to "feature" weighting.

This improved reliability does not come for free, and more computational time is required. For instance, Punch *et al.* [14] reported experiments with genetic algorithms for feature selection that took 14 days. In this case, one might not want genetic algorithms, since the difference in accuracy when compared with other approaches may not be great enough to justify the extra effort.

Alternatively, one could just gather more labelled data. For example, the U.S. Geological Survey has produced a data set of about 400 labelled rocks. However, some of these labels are wrong, or inconsistent with the classification scheme of the JPL data set. Before combining these data with the JHU data, additional preprocessing would be required. Also, The Arizona State University Spectral Library contains a sizeable and well-characterized collection of mineral spectra in the 0.6 - 25.0 μm (1600 - 400 cm^{-1}) range [1], measured using a Thermal Emission Spectrometer (TES). This is significant because of the interest of TES spectroscopy to the Mars program. Once a similarly sized and well-characterized library of rock TES spectra becomes available in overlapping range, it should be possible to generate masks for analyzing data in the infrared range that will have direct application to analyzing the voluminous databases of TES spectra that have been measured by orbital and surface-based instruments of Martian rocks and soils.

Concerning the variability of the underlying classifier, a straightforward way to alleviate this problem is to modify the evaluation function of the search algorithms to consider the outcome of an ensemble of classifiers. Future experiments may include this approach.

8 Acknowledgements

This work was partially funded by the National Aeronautics & Space Administration, grant number NAG5-9309. Jonathan Moody was partially supported by a National Science Foundation Graduate Research Fellowship. Ricardo Silva was partially supported by a Microsoft Fellowship.

References

[1] Christensen, P.; Bandfield, J.; Hamilton, V.; Howard, D.; Lane, M.; Piatek, J.; Ruff, S. & Stefanov, W. (2000). A thermal emission spectral library of rock-forming minerals. J. Geophys. Res., 105, 9735-9739.

[2] Cohen, P. (1995). Empirical Methods for Artificial Intelligence. MIT Press.

[3] Cohen, P. and Jensen, D. (1997). "Overfitting Explained". Preliminary Papers of the Sixth International Workshop on Artificial Intelligence and Statistics, 115-122.

[4] Demiroz, G. & Guvenir, H. (1996). "Genetic algorithms to learn feature weights for the nearest neighbor algorithm". In Proceedings of BENELEARN-96, 117-126.

[5] Friedman, J. & Fisher, N. (1999). "Bump hunting in high-dimensional data". Statistics and Computing, 9, 123-143.

[6] Goldberg, D. (1989). Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley.

[7] Guyon, I.; Matic, N. & Vapnik, V. (1995). "Discovering Informative Patterns and Data Cleaning". In: Advances in Knowledge Discovery and Data Mining, 181-204. AAAI Press.

[8] Hall, M. (1999). Correlation-based Feature Selection for Machine Learning. PhD thesis, University of Waikato, Computer Science Department. Hamilton, New Zealand.

[9] Hapke, B. (1993). Theory of Reflectance and Emittance Spectroscopy. Cambridge University Press, New York.

[10] Kohavi, R. and John, G. (1998) "The Wrapper Approach". In H. Liu and H. Motoda (Eds.), Feature Selection for Knowledge Discovery in Databases. Springer-Verlag.

[11] Masters, T. (1993). Neural Network Recipes in C++. Academic Press.

[12] Merenyi, E. (2000) "Precision Mining" of High-Dimensional Patterns with Self-Organizing Maps: Interpretation of Hyperspectral Images. In Quo Vadis Computational Intelligence: New Trends and Approaches in Computational Intelligence, Studies in Fuzziness and Soft Computing 54.

[13] Pieters, C. and P. Englert, editors (1993). Remote Geochemical Analysis: Elemental and Mineralogical composition. Cambridge University Press, New York.

[14] Punch, W.; Goodman, E.; Pei, M.; Chia-Shun, L.; Hovland, P. & R. Enbody (1993). "Further research on feature selection and classification using genetic algorithms". Proceedings of the International Conference on Genetic Algorithms 93, 557-564.

[15] Pyle, D. (1999). Data Preparation for Data Mining. Morgan-Kaufmann.

[16] Ramsey, Joseph (2001) Expertise and Mixture in Automatic Causal Discovery. PhD Thesis, Dept. of Philosophy, University of California, San Diego.

[17] Ramsey, Joseph; Gazis, Paul; Roush, Ted; Spirtes, Peter; Glymour, Clark. (2002). "Automated Remote Sensing with Near Infrared Reflectance Spectra: Carbonate Recognition". Data Mining and Knowledge Discovery Journal (to appear).

[18] Salisbury, J.; Walter, L. & D'Aria, D. (1988) Mid-Infrared (2.5-13.5 micrometer) Spectra of Igneous Rocks: Open-File Report 88-686.

[19] Salisbury, J.; Walter, L.; Vergo, N. & D'Aria, D. (1991). Infrared (2.1-25 um) Spectra of Minerals. Baltimore: JohnsHopkins University Press.

[20] Sarle, W. (2000). The Neural Network FAQ.ftp://ftp.sas.com/pub/neural/FAQ.html.

[21] Spirtes, P., Glymour, C., & Scheines, R. (2000). Causation, Prediction and Search, 2nd edition. MIT Press.

[22] Vafaie, V. & DeJong, K. (1998). "Feature space transformation using genetic algorithms". IEEE Transactions on Intelligent Systems, 13(2), 57-65.

[23] Wettschereck, D., Aha, D. W., & Mohri, T. (1997). "A review and comparative evaluation of feature weighting methods for lazy learning algorithms". Artificial Intelligence Review, 11, 273-31.

[24] Yang, J. & Honavar, V. (1998). "Feature subset selection using a genetic algorithm". In Feature Extraction, Construction, and Subset Selection: A Data Mining Perspective. Motoda, H. and Liu, H. (Ed.) New York: Kluwer.