RARE JAROSITE DETECTION IN CRISM IMAGERY BY NON-PARAMETRIC BAYESIAN CLUSTERING

Murat Dundar *
Indiana University - Purdue University
Computer & Information Science Department
Indianapolis, Indiana, 46202 USA

Bethany L. Ehlmann
California Institute of Technology
Division of Geological & Planetary Sciences
Jet Propulsion Laboratory
Pasadena, California, 91125 USA

ABSTRACT

Discovery of rare phases on Mars is important as they serve as indicators of the geochemistry of the Mars surface and facilitate understanding of mineral assemblages within a geologic unit. Identification of rare minerals in high spatial and spectral resolution Compact Reconnaissance Imaging Spectrometer for Mars (CRISM) visible/shortwave infrared (VSWIR) images has been a challenge due to the presence of both additive and multiplicative noise and other artifacts, affecting all collected images, in addition to the limited spatial extent of regions hosting these minerals. In an effort to automate this task we evaluate various clustering algorithms using the detection of rare jarosite, associated with spectrally similar minerals in CRISM imagery, as a case study. We compare non-parametric Bayesian and standard clustering algorithms and show that a recently developed doubly non-parametric Bayesian model could be effective for this task.

Index Terms— CRISM, jarosite, rare target detection, non-parametric bayesian, clustering

1. INTRODUCTION

Discovery of small, rare phases on Mars is important for two reasons. First, specific minerals such as alunite and jarosite (acidic), serpentine (alkaline, reducing), analcime (alkaline, saline), prehnite (200°C - 400°C), and perhaps phases yet to be discovered serve as direct environmental indicators of the chemistry of waters on the Mars surface. Second, the identification of rare endmember phases facilitates understanding the mineral assemblages within a geologic unit, which are critical for identifying the thermodynamic conditions and fluid composition during interactions of rocks with liquid water.

The identification of these uncommon and spatially restricted mineral phases is difficult using existing CRISM processing techniques. The most common spectral mineral-identification method involves finding the ratio of the average spectra from two regions along-track in the image, where the numerator is the spectrum from the area of interest and the denominator is the spectrum derived from a spectrally homogeneous “neutral” region [1]. Spatial averaging filters additive random noise, whereas spectral ratioing reduces multiplicative noise. Summary parameters [2, 3] derived from key absorption bands are used to identify candidate regions for the numerator and denominator. Ratioing has proven very effective in the identification of minerals that occur over relatively large spatial extents in images. However, summary parameters have had limited success to date for identification of rare phases spanning a limited number of contiguous pixels in an image.

We consider rare jarosite detection as a case study and investigate the performances of several clustering algorithms toward automating rare phase detection in CRISM imagery. Jarosite is important because it is an indicator mineral for acidic, oxidizing conditions. Jarosite can be challenging to detect because its principal absorption occurs in a spectral region shared by absorptions in aluminum phyllosilicates and silica phases (Fig. 1) as well as iron-rich phyllosilicates, common mineral classes on Mars, which form spatially extensive geologic units [4].

2. METHODS

2.1. Image preprocessing and ground truth

Two images from Nili Fossae (FRT00009971, FRT0000A053) and one image from Mawrth Vallis (HRL000043EC) were used. Simple atmospheric and photometric corrections are applied to all three images using the CRISM Analysis Toolkit [7, 8]. Only the spectral channels that cover the spectral region from 1.0 to 2.6 μm (248 channels) are used in this study. The channels corresponding to the remaining part of the spectrum (0.4-1.0 μm and 2.6 to 4.0 μm) were excluded because surface spectral properties at shorter wavelengths are
Jarosite locations are initially identified using the I²GMM algorithm [9]. Because this algorithm is one of the algorithms used in this study for comparison, to avoid bias in establishing our ground truth, the final spatial extent of jarosite in each image is determined by visual inspection by investigating several ratioed spectra around this location and choosing the region that best recovers the distinctive doublet absorption in the 2.20-2.27 µm region characteristic of jarosite in CRISM imagery. For all three images the extracted spectra are shown (Fig. 2).

Clustering experiments are performed with approximately 10K pixels obtained from 100 by 100 sub images including ground truth jarosite locations. Sub-images rather than original images are used for two reasons. First, for each algorithm several different parameter configurations are considered and for each configuration experiments are run ten times to measure variations in performance measure across different runs. Using a sub-image significantly reduces run-time for each algorithm. Second, for more accurate evaluation of the clustering algorithms, it is important to verify that all false positives identified by each algorithm are indeed false positives and not other jarosite regions that were overlooked when establishing the ground truth. This task can be more easily performed for a sub-image than the original larger image. Each pixel is projected onto the top ten principal components of the image I/F data. I/F data is derived from the radiance data by computing the ratio of the radiance to the solar irradiance at Mars [13].

Data points in the following presentation refers to pixels represented as points in this 10-dimensional vector space.

2.2. Standard clustering algorithms

The K-means algorithm and finite Gaussian mixture model (GMM) are popular algorithms for clustering data sets. K-means fixes the number of clusters K beforehand and assigns each pixel to one of the K clusters by minimizing a predefined distance metric in an iterative fashion. Similar to the K-means, finite GMM also fixes the number of clusters K beforehand, but unlike K-means fits a K-component Gaussian mixture model onto the data using the expectation-maximization algorithm [14]. K-means can be considered as a special case of GMM that restricts component covariances to spherical shapes.

2.3. Non-parametric Bayesian clustering algorithms

Finite GMM and K-means algorithms have two major limitations. First, no prior knowledge about cluster characteristics are used during clustering. Although lack of prior knowledge may not pose a serious problem for clustering data sets with balanced and well-defined clusters, prior knowledge may become critical when clustering data sets with rare clusters. Second, both finite GMM and K-means require that the number of components is defined in advance. Although there are several ways to predict the number of components in the data in an offline manner, these techniques are in general suboptimal as they decouple the two interdependent tasks: predicting the number of components and predicting model parameters.

A more flexible version of GMM can be derived by taking the limit over the number of mixture components to infinity. With this infinite version of GMM (IGMM) the actual number of components is automatically estimated during inference along with other component parameters in a single unified process [15]. IGMM is considered a non-parametric model as the number of components is no longer fixed during inference and can arbitrarily grow to better accommodate data sets as needed. It is also Bayesian as prior knowledge about number of clusters, cluster shapes and dispersion can be encoded into the model. [16].

\[
\begin{align*}
x_i & \sim P(x_i | \theta_i) \\
\theta_i & \sim G \\
G & \sim DP(\alpha H) \\
H & = N(\mu_1, \Sigma_{10}^{-1})W^{-1}(\Sigma|\Sigma_0, m)
\end{align*}
\]

The IGMM model used in this study is shown in (1). According to this model individual data points \(x_i \in R^d\) are gen-
Fig. 2. (A) Averaged spectra for ground truth jarosite locations. Jarosite is identified primarily by the 2.26-2.27 \( \mu \text{m} \) absorption feature, though other absorptions are sometimes present. Library spectra from [5] and [10] are shown for comparison. The CRISM spectrum from image FRT000043EC is jarosite from an area in Mawrth Vallis previously reported in [11] to contain ferricopiapite, another indicator of acid sulfate alteration. FRT00009971 and FRT0000A053 are new detections from the Nili Fossae region. The context of FRT00009971 was previously described in [12]. (B) Follow-on ‘‘expert user” analyses based on findings from FRT0000A053 confirm that jarosite-bearing terrains possess discrete spectral properties relative to other scene materials and permit spatial mapping of the phase in geologic context, here in association with an aluminum phyllosilicate-bearing stratigraphic unit. Colors correspond to the minerals indicated, with the figure generated by overlaying mapping of band strength at 2.24-2.27 \( \mu \text{m} \) (red); 2.28-2.31 \( \mu \text{m} \) (green) and 2.18-2.22 \( \mu \text{m} \) (blue) on a CRISM infrared albedo map. Library spectra from [5] are compared to spectra from regions in the CRISM image.

erated from Gaussian distributions whose parameters are in turn drawn from a Dirichlet Process (DP). A Dirichlet Process (DP) is a distribution over distributions. It generates random distribution \( G \) based on parameters \( H \) and \( \alpha \). The parameter \( H \) defines the base distribution from which the discrete probability masses, i.e., atoms, of \( G \) are drawn. The concentration parameter \( \alpha \) changes the sparseness of \( G \), which in turns affects the number of observed components in mixture models. The random measure \( G \) can be considered as a mixture of infinitely many atoms with their mixture weights drawn from a stick-breaking distribution.

As data points are modeled by Gaussian clusters the base distribution \( H \) of DP serves as a prior over the cluster mean vectors \( \mu_i \) and covariance matrices \( \Sigma_i \) in which case \( \theta_i = \{ \mu_i, \Sigma_i \} \). In this case a conjugate distribution over a Gaussian data model would be a bivariate prior \( H \) that involves a Gaussian prior for the mean vectors and Inverse Wishart prior for the covariance matrices as in (1). We will denote this bivariate distribution by NIW.

NIW includes four hyperparameters: \( \{ \mu_0, \kappa_0, \Sigma_0, m \} \). The hyperparameter \( \mu_0 \) is the mean of the Gaussian prior defined over the cluster means. The hyperparameter \( \kappa_0 \) is a scaling constant that adjusts the dispersion of the cluster centers around \( \mu_0 \). A smaller value for \( \kappa_0 \) suggests that cluster centers are expected to be farther apart from each other whereas a larger value suggests cluster centers closer to each other. The hyperparameters \( \Sigma_0 \) and \( m \) dictate the expected shape of the clusters. The minimum feasible value of \( m \) is equal to \( d + 2 \), and the larger the \( m \), the less individual covariance matrices will deviate from the expected shape.

Both IGMM and GMM models each cluster with a single Gaussian component. This is a serious limitation if clusters emerge with multi-mode and/or skewed distributions, in which case IGMM creates additional components to better fit the data set. As there is no hierarchy in IGMM to allow for grouping of components into clusters, these additional components are treated as separate clusters, leading to suboptimal clustering performance. The infinite mixtures of infinite Gaussian mixture model (\( \infty^2 \)GMM) is developed to circumvent this limitation by offering a two-layer non-parametric GMM [9]. This model is doubly non-parametric in terms of the number of clusters and the number of components for each
cluster, allowing for modeling clusters of various shapes. The generative model of IGMM is given in (2).

\[
H = NIW(\mu_0, \Sigma_0, \kappa_0, m) \\
G \sim DP(\gamma H) \\
\theta_j = (\mu_j, \Sigma_j) \sim G \\
H_j = N(\mu_j, \Sigma_j/\kappa_1) \\
G_j \sim DP(\alpha H) \\
\mu_{ji} \sim G_j \\
x_{ji} \sim N(\mu_{ji}, \Sigma_j)
\]  

(2)

According to this generative model a discrete mixing measure \( G \) is sampled from a global Dirichlet Process with base distribution \( H \) and concentration parameter \( \gamma \). Cluster centers \( \mu_j \) and covariances \( \Sigma_j \) are drawn from \( G \). Local DPs are defined, one for each cluster generated by the global DP. The base distributions \( H_j \) of local DPs are Gaussian distributions centered at \( \mu_j \) with a covariance of \( \kappa_1^{-1} \Sigma_j \).

Cluster-specific discrete mixing measures \( G_j \) are drawn from their corresponding local DPs with base distribution \( H_j \) and concentration parameter \( \alpha \). The mean vectors \( \mu_{ji} \) of components associated with cluster \( j \) are drawn from \( G_j \). Data points are generated from Gaussian components with mean vector \( \mu_{ji} \) and covariance matrix \( \Sigma_j \). In a nutshell this hierarchical model creates two level partitioning by clustering data points into components and components into clusters.

In addition to hyperparameters that already exist in the IGMM model, IGGM introduces two additional hyperparameters \( \kappa_1 \) and \( \gamma \). The hyperparameter \( \kappa_1 \) models dispersion of component means around corresponding cluster center. A larger \( \kappa_1 \) leads to smaller variations in component means with respect to the corresponding cluster center, generating Gaussian-like clusters, whereas a smaller \( \kappa_1 \) dictates larger variations generating more flexible cluster shapes. The hyperparameter \( \gamma \) adjusts the number of clusters generated while \( \alpha \) controls the number of components generated for each cluster.

\[m, \kappa_0, \kappa_1, s\) to tune for IGGM for which thirty six quartets are considered. These values are chosen to cover a broad range of cluster characteristics as discussed in [17]. For each clustering configuration \( F_1 \) scores are computed for all clusters and the cluster with the highest \( F_1 \) is considered as the most promising jarosite detection by each technique.

\[
\begin{array}{cccc}
\text{K-means} & \text{GMM} & \text{IGMM} & \text{IGGM} \\
9971 & 0.60/450 & 0.59/450 & 0.48/392 & 0.77/45 \\
4053 & 0.46/300 & 0.44/500 & 0.46/195 & 0.70/34 \\
43EC & 0.86/400 & 0.63/400 & 0.78/190 & 0.83/101
\end{array}
\]

Table 1. \( F_1 \) scores and number of clusters \( K \) provided in the format \( F_1/K \). Results are averages over ten runs.

The results in Table 1 favors IGGM as a promising approach for rare jarosite detection in terms of both accuracy measured by \( F_1 \) score and efficiency measured by the number of clusters generated. Despite generating a significantly less number of clusters than other techniques, IGGM achieves the highest \( F_1 \) score for two of the images and produces an \( F_1 \) score that is comparable to K-means for the third image.

4. CONCLUSIONS

Automated rare phase detection in CRISM imagery is an important albeit a challenging problem that requires flexible models. A mixture model with a large number of components can more accurately estimate the density of the data. However, more accurate estimation of the density function may not improve clustering accuracy as the problem of many-to-one mappings between components and clusters remains unsolved in standard mixture models. IGGM addresses this limitation by jointly clustering data points into components and components into clusters in a unified model inference offering extreme flexibility in modeling a wide array of data distributions. Our preliminary results demonstrate the promising aspect of this framework for rare jarosite detection. Hyperparameter tuning is not specifically addressed in this study as we are still in the early stages of developing domain knowledge. An optimal parameter set that can more effectively encode domain knowledge can be obtained by processing several hundred images from different regions of Mars and performing both qualitative and quantitative evaluation of different hyperparameter sets. Additional details about technical aspects of IGGM and its implementation can be found in [17].
5. REFERENCES


