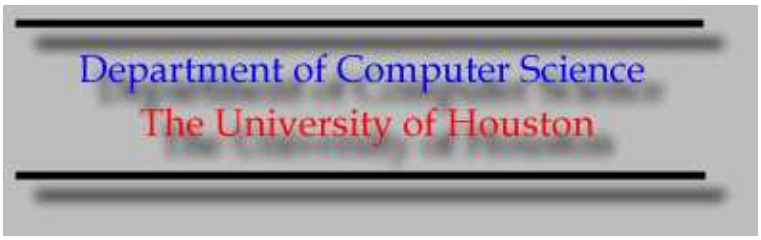

An Efficient Approach to External Cluster Assessment with an Application to Martian Topography*

Ricardo Vilalta, Tom Stepinski†, and Muralikrishna Achari

Department of Computer Science
University of Houston
Houston, TX, 77204, USA
http://www.cs.uh.edu

Technical Report Number UH-CS-05-08

April 20, 2005

Keywords: external cluster validation, multivariate Gaussian distributions, Martian topography.

Abstract

Automated tools for knowledge discovery are frequently invoked in databases where objects already group into some known (i.e., external) classification scheme. In the context of unsupervised learning or clustering, such tools delve inside large databases looking for alternative classification schemes that are meaningful and novel. An assessment of the information gained with new clusters can be effected by looking at the degree of separation between each new cluster and its most similar class. Our approach models each cluster and class as a multivariate Gaussian distribution and estimates their degree of separation through an information theoretic measure (i.e., through relative entropy or Kullback Leibler distance). The inherently large computational cost of this step is alleviated by first projecting all data over the single dimension that best separates both distributions (using Fisher's Linear Discriminant). We test our algorithm on a dataset of Martian surfaces using the traditional division into geological units as external classes and the new, hydrology-inspired, automatically performed division as novel clusters. We find the new partitioning constitutes a formally meaningful classification that deviates substantially from the traditional classification.

UNIVERSITY of HOUSTON

*This material is based upon work supported by the National Science Foundation under Grants no. IIS-0431130, IIS-0448542, and IIS-0430208.

†T. Stepinski is with the Lunar and Planetary Institute, 3600 Bay Area Blvd, Houston TX 77058-1113, USA.

An Efficient Approach to External Cluster Assessment with an Application to Martian Topography*

Ricardo Vilalta, Tom Stepinski†, and Muralikrishna Achari

Abstract

Automated tools for knowledge discovery are frequently invoked in databases where objects already group into some known (i.e., external) classification scheme. In the context of unsupervised learning or clustering, such tools delve inside large databases looking for alternative classification schemes that are meaningful and novel. An assessment of the information gained with new clusters can be effected by looking at the degree of separation between each new cluster and its most similar class. Our approach models each cluster and class as a multivariate Gaussian distribution and estimates their degree of separation through an information theoretic measure (i.e., through relative entropy or Kullback Leibler distance). The inherently large computational cost of this step is alleviated by first projecting all data over the single dimension that best separates both distributions (using Fisher's Linear Discriminant). We test our algorithm on a dataset of Martian surfaces using the traditional division into geological units as external classes and the new, hydrology-inspired, automatically performed division as novel clusters. We find the new partitioning constitutes a formally meaningful classification that deviates substantially from the traditional classification.

Index Terms

external cluster validation, multivariate Gaussian distributions, Martian topography.

I. Introduction

Clustering algorithms are useful tools in revealing structure from unlabeled data; the goal is to discover how data objects gather into natural groups. Research spans multiple topics such as the cluster representation (e.g., flat, hierarchical), the criterion function to identify sensible clusters (e.g., sum-of-squared errors, minimum variance), and the proximity measure that quantifies the degree of similarity (conversely dissimilarity) between data objects (e.g, Euclidean distance, Manhattan norm, inner product). Additionally, the application of clustering algorithms can be preceded and followed by various steps. First, cluster tendency is a preprocessing step that indicates when data objects exhibit a clustering structure; it precludes using clustering when the data appears randomly generated under the uniform distribution over a sample window of interest in the attribute space [1], [2], [3], [4], [5]. Second, cluster validation is a postprocessing step that is most necessary to assess the quality and meaning of the resulting clusters [6], [7], [8].

Cluster validation plays a key role in assessing the value of the output of a clustering algorithm by computing statistics over the clustering structure. Cluster validation is called *internal* when statistics are devised to capture the quality of the induced clusters using the available data objects only [9], [10], [8]. If the validation is performed by gathering statistics comparing the induced clusters against an external and independent classification of objects, the validation is called *external*[3]. External cluster validation is based on the assumption that an understanding of the output of the clustering algorithm can be achieved by finding a resemblance of the clusters with existing classes [11], [12], [6], [7], [13]. Such narrow assumption precludes alternative interpretations; in some scenarios high-quality clusters (as supported by an internal validation step) are considered novel if they do not resemble existing classes. We prefer to employ the term *external cluster assessment* when referring to a methodology intended to quantify the value of new clusters when compared to an external and independent classification scheme. This adds flexibility to the validation task. In some scenarios, a large separation between clusters and classes serves to indicate cluster

*This material is based upon work supported by the National Science Foundation under Grants no. IIS-0431130, IIS-0448542, and IIS-0430208.

‡T. Stepinski is with the Lunar and Planetary Institute, 3600 Bay Area Blvd, Houston TX 77058-1113, USA.

novelty [14]; on the other hand, finding clusters resembling existing classes serves to confirm existing theories of data distributions [8]. Both types of interpretations are legitimate; the value of new clusters is ultimately decided by domain experts after careful interpretation of the distribution of new clusters and existing classes.

In this paper we propose a method for external cluster assessment that runs contrary to the traditional view of external cluster validation; most traditional metrics output a single value indicating the degree of match between the partition induced by the known classes and the one induced by the clusters. Our goal instead is to compute the distance between each individual cluster and its most similar external class; our method works efficiently by projecting the data to a single dimension that best captures the true separation between the class-cluster pair on the original attribute space. Traditional metrics cannot be easily compared to our approach for several reasons. First, by averaging the degree of match across all classes and clusters, such metrics fail to identify the potential value of individual clusters. Moreover, the lack of a probabilistic model in the representation of data distributions precludes projecting the extent to which a class-cluster pair intersect. Our approach differs in using a probabilistic model to evaluate each cluster individually, ranking all classes against each cluster based on their degree of overlap or intersection.

We apply our methodology on the characterization and classification of surfaces on Mars. The planet Mars is at the center of our solar system exploration efforts. There are several Mars orbiters remotely collecting imagery, topographic, and spectral data of the planet's surface. The current principal tool for studying Martian surfaces is geologic mapping. The standard technique of photogeologic interpretation of images [15] has been developed to facilitate such mapping. A collection of sites on Mars constitutes a set of objects that are classified manually by domain experts (geologists) on the basis of their geological properties. The resultant division of sites into the so-called "geological units" (see section 4.1) represents an external classification. Geologic mapping, however, is a slow and subjective procedure. The availability of Martian digital topography data suggests an alternative classification of Martian sites based exclusively on selected topographical properties. Specifically, a relatively simple mathematical representation [16], [17] can be constructed for a site's "drainage" network (see section 4.2). A quantitative representation enables an automated, objective, and fast comparison between different sites. We have constructed such representation for a large set of Martian sites and have used a clustering algorithm to divide this set into natural groups. Using our approach to external cluster assessment, we study whether this novel partitioning resembles the traditional, external classification. We find the new partitioning, based on hydrology-inspired variables, deviates substantially from the traditional classification.

This paper is organized as follows. Section II provides background information and defines traditional metrics for external cluster validation. Section III explains our proposed metric. Section IV describes our domain of study based on the characterization of Martian surfaces. Section V reports on the results of clustering Martian sites on the basis of their topographic properties, and provides an interpretation of the output clusters. Lastly, section VI gives our summary and discusses future work.

II. Preliminaries: External Cluster Validation

We assume a dataset of objects, $\mathcal{D} : \{\mathbf{x}_i\}$, where each $\mathbf{x}_i = (a_1, a_2, \cdots, a_k)$ is an attribute vector characterizing a particular object. We refer to an attribute variable as A_i, and to a particular value of that variable as a_i. The space $\mathcal{X}$ of all possible attribute vectors is called the attribute space. We make the simplifying assumption that each attribute value is a real number, and thus $\mathbf{x}_i \in \Re^k$.

A clustering algorithm partitions $\mathcal{D}$ into n mutually exclusive and exhaustive[4] subsets $\mathcal{K}_1, \mathcal{K}_2, \cdots, \mathcal{K}_n$, where $\bigcup_j \mathcal{K}_j = \mathcal{D}$. Each subset $\mathcal{K}_j$ represents a cluster. The goal of a clustering algorithm is to partition the data such that the average distance between objects in the same cluster (i.e., the average intra-distance) is significantly less than the distance between objects in different clusters (i.e., the average inter-distance) [18]. Distances are measured according to some predefined metric (e.g., Euclidean distance, Manhattan norm, inner product) over space $\mathcal{X}$.

We assume the existence of a different mutually exclusive and exhaustive partition of objects, $\mathcal{C}_1$, $\mathcal{C}_2, \cdots$, $\mathcal{C}_m$, where $\bigcup_i \mathcal{C}_i = \mathcal{D}$, induced by a natural classification scheme that is independent of the partition induced by the clustering algorithm. Note that the number of external classes need not match the number of clusters. Our goal is to perform an objective comparison of both partitions. It must be emphasized that the previously known classification is independent of the induced clusters since our main goal is to ascribe a meaning to the partition induced by the clustering algorithm; one may even use multiple existing external classifications to validate the set of induced clusters.

Traditionally, the goal behind external cluster validation is to find a near-optimal match between clusters and external classes; if found we say the clusters have simply recovered the external class structure. As mentioned above, we suggest a broader goal where a form of external cluster assessment indicates the degree of separation between clusters and classes; the scientific value behind a near match or strong disagreement can then be elicited through domain expertise.

A. Metrics Comparing Classes and Clusters

In this section we briefly review representative work in the field of external cluster validation. Several approaches exist attacking the problem of assessing the degree of match between the set $\mathcal{C} = \{\mathcal{C}_i\}$ of predefined classes and the set $\mathcal{K} = \{\mathcal{K}_j\}$ of new clusters. In all cases high values indicate a high similarity between classes and clusters. We divide these approaches based on the kind of statistics employed.

The 2×2 Contingency Table

One type of statistical metrics is defined in terms of a 2×2 table where each entry $\mathcal{E}_{ij}$, $i, j \in \{1, 2\}$, counts the number of object pairs that agree or disagree with the class and cluster to which they belong; $\mathcal{E}_{11}$ corresponds to the number of object pairs that belong to the same class and cluster, similar definitions apply to other entries where $\mathcal{E}_{12}$ corresponds to same class and different cluster, $\mathcal{E}_{21}$ corresponds to different class and same cluster, and $\mathcal{E}_{22}$ corresponds to different class and different cluster. Clearly $\mathcal{E}_{11}$ and $\mathcal{E}_{22}$ denote the number of object pairs contributing to a high similarity between classes and clusters, whereas $\mathcal{E}_{12}$ and $\mathcal{E}_{21}$ denote the number of object pairs contributing to a high degree of dissimilarity. Let P be the total number of possible object pairs (if N is the total number of data objects, then $P = \frac{N(N-1)}{2}$). The following statistics have been suggested as metrics of similarity or overlap:

Rand [6]:

$$\frac{\mathcal{E}_{11} + \mathcal{E}_{22}}{\mathcal{E}_{11} + \mathcal{E}_{12} + \mathcal{E}_{21} + \mathcal{E}_{22}} \tag{1}$$

Jaccard [13]:

$$\frac{\mathcal{E}_{11}}{\mathcal{E}_{11} + \mathcal{E}_{12} + \mathcal{E}_{21}} \tag{2}$$

Fowlkes and Mallows [7]:

$$\frac{\mathcal{E}_{11}}{\sqrt{(\mathcal{E}_{11} + \mathcal{E}_{12})(\mathcal{E}_{11} + \mathcal{E}_{21})}} \tag{3}$$

Γ statistic [19]:

$$\frac{P\mathcal{E}_{11} - (\mathcal{E}_{11} + \mathcal{E}_{21})(\mathcal{E}_{11} + \mathcal{E}_{21})}{\sqrt{(\mathcal{E}_{11} + \mathcal{E}_{21})(\mathcal{E}_{11} + \mathcal{E}_{21})(P - (\mathcal{E}_{11} + \mathcal{E}_{21}))(P - (\mathcal{E}_{11} + \mathcal{E}_{21}))}} \tag{4}$$

Experiments using artificial datasets show these metrics have good convergence properties (i.e., converge to maximum similarity if classes and clusters are identically distributed) as the number of clusters and dimensionality increase [13].

The $m \times n$ Contingency Table

A different approach is to work on a contingency table $\mathcal{M}$, defined as a matrix of size $m \times n$ where each row correspond to an external class and each column to a cluster. An entry $\mathcal{M}_{ij}$ indicates the number of objects covered by class $\mathcal{C}_i$ and cluster $\mathcal{K}_j$.

Using $\mathcal{M}$, the simmilarity between $\mathcal{C}$ and $\mathcal{K}$ can be defined in several forms:

Normalized Hamming Distance [12]:

$$\frac{DH_c(\mathcal{M}) + DH_k(\mathcal{M})}{2N} \tag{5}$$

where $N = |\mathcal{D}|$ is the size of the dataset (i.e., where $N = \sum_i \sum_j \mathcal{M}_{ij}$) and the directional Hamming distances are defined as follows:

$$DH_c(\mathcal{M}) = \sum_i \max_j \mathcal{M}_{ij} \tag{6}$$

$$DH_k(\mathcal{M}) = \sum_j \max_i \mathcal{M}_{ij} \tag{7}$$

Equation 5 measures accuracy by adding the highest value on each row (conversely column) in $\mathcal{M}$ divided by the total number of objects. Rows and columns are worked out separately since the number of classes and clusters may be different.

Empirical Conditional Entropy [11], [20]:

$$H(C|K) = -\sum_i \sum_j \frac{\mathcal{M}_{ij}}{N} \log_2 \frac{\mathcal{M}_{ij}}{\mathcal{M}_j} \tag{8}$$

where $\mathcal{M}_j$ is the marginal sum $\sum_i \mathcal{M}_{ij}$ and lower values are preferred. Equation 8 measures the degree of impurity of the partitions induced by the clustering algorithm and is biased towards distributions characterized by many clusters; this bias can be adjusted by applying the minimum description length principle [11].

Hypothesis Testing

The metrics described above serve as an indication of the proximity between the set $\mathcal{C} = \{\mathcal{C}_i\}$ of classes and the set $\mathcal{K} = \{\mathcal{K}_j\}$ of newly constructed clusters. In addition one must be able to show that the output metric score (i.e., output statistic) is far from the value one would obtain if the structure of $\mathcal{D}$ were random (if the objects $\mathcal{D}$ were uniformly distributed). To decide if the null hypothesis H_0 (that assumes a uniform distribution over $\mathcal{D}$) can be rejected, we rely on Monte-Carlo techniques [21], [22]. The output statistic s is compared against the set $\{s_i\}_{i=1}^r$ of statistics gathered assuming the null hypothesis true. This is effected by applying the clustering algorithm to r different artificial samples where data objects distribute uniformly randomly; on each sample we compute statistic s_i. The null hypothesis H_0 is rejected if s is greater than$(1-\rho)r$ of the s_i values (for a given significance level ρ).

B. Limitations of Current Metrics

In practice, a quantification of the similarity between sets of classes and clusters is of limited value; any potential discovery provided by the clustering algorithm is only identifiable by analyzing the meaning of each cluster individually. As an illustration, Figure 1(left) shows a two-dimensional attribute space where two clusters ($\mathcal{K}_1$, $\mathcal{K}_2$) make a close match with two external classes ($\mathcal{C}_1$, $\mathcal{C}_2$) ; a third cluster ($\mathcal{K}_3$) denotes a novel structure that does not resemble any existing classes. Averaging the similarity between clusters and classes altogether disregards the potential discovery carried by the third cluster.

In addition, even when in principle one could analyze the entries of a contingency matrix to identify clusters having little overlap with existing classes (section II-A), such information cannot be used in estimating the intersection of the true probability models from which the objects are drawn. This is because the lack of a probabilistic model in the representation of data distributions precludes estimating the extent of the intersection of a class-cluster pair. As an illustration, Figure 1(right) shows a two-dimensional attribute space comparing a cluster $\mathcal{K}_j$ with an external class $\mathcal{C}_i$. The z axis represents the conditional probability of a data object ($P(\mathbf{x}|\mathcal{K}_j)$ for cluster $\mathcal{K}_j$ and $P(\mathbf{x}|\mathcal{C}_i)$ for class $\mathcal{C}_i$). A contingency matrix simply counts the number of data objects falling on different regions of the attribute space (e.g., $\mathcal{K}_j \cap \mathcal{C}_i$, $\mathcal{K}_j \setminus \mathcal{C}_i$, $\mathcal{C}_i \setminus \mathcal{K}_j$, $\overline{\mathcal{K}_j \cup \mathcal{C}_i}$); a probabilistic model, in contrast, generates an expectation of the number of objects lying on these regions; the expectation can differ significantly from the actual count. This is the result of constructing density models using all data objects that belong to the class-cluster pair of interest. We address these issues and our proposed metric next.

Fig. 1. (left) Averaging the similarity between clusters and classes altogether disregards the potential discovery carried by cluster $\mathcal{K}_3$. (right) A contingency matrix simply counts the number of objects covered by both class and cluster; a probabilistic model generates an expectation based on the density of that intersection that may differ significantly from the actual count.

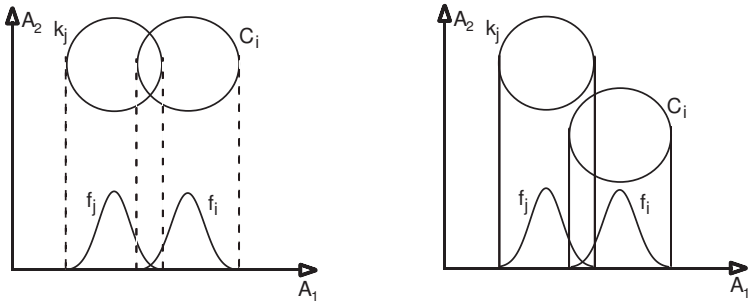

Fig. 2. (left) A projection of the data over an attribute transforms the original problem into a new problem made of one-dimensional Gaussian distributions; (right) Two non-overlapping distributions in a k-dimensional space may appear highly overlapped when projected over each attribute (here $k = 2$).

III. Our Approach to External Cluster Assessment

We now turn into our proposed approach for external cluster assessment. We start under the assumption that both clusters and classes can be modelled using a multi-variate Gaussian (i.e., Normal) distribution. In this case the probability density function is completely defined by a mean vector μ and covariance matrix $\mathbf{\Sigma}$:

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{k/2}|\mathbf{\Sigma}|^{1/2}} \exp\ [-\frac{1}{2}(\mathbf{x}-\mu)^t\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)] \tag{9}$$

where $\mathbf{x}$ and μ are k-component vectors, and $|\mathbf{\Sigma}|$ and $\mathbf{\Sigma}^{-1}$ are the determinant and inverse of the covariance matrix.

Our goal is simply to assess the separation or distance between a cluster $\mathcal{K}_j$, modelled as $f_j(\mathbf{x}) : N[\mu_j, \mathbf{\Sigma}_j]$, and its most similar class $\mathcal{C}_i$, modelled as $f_i(\mathbf{x}) : N[\mu_i, \mathbf{\Sigma}_i]$ (where f_j corresponds to $P(\mathbf{x}|\mathcal{K}_j)$ and f_i corresponds to $P(\mathbf{x}|\mathcal{C}_i)$). Before explaining our methodology (section III-C) we introduce two preliminary metrics.

A. Integrating Over the Attribute Space

A straightforward approach to measure the degree of separation between $f_j(\mathbf{x})$ and $f_i(\mathbf{x})$, denoted as $\Psi(f_j, f_i)$, is to apply a function to each point $\mathbf{x}$, $\psi_{f_j,f_i}(\mathbf{x})$, and to integrate that function over the whole attribute space:

$$\Psi(f_j, f_i) = \int_{\mathbf{x}} \psi_{f_j,f_i}(\mathbf{x})\ d\mathbf{x} \tag{10}$$

A simple example of $\psi_{f_j,f_i}(\mathbf{x})$ is the square distance $(f_j(\mathbf{x}) - f_i(\mathbf{x}))^2$. We assume that in the extreme case where both distributions are identical then $\psi_{f_j,f_i}(\mathbf{x}) = 0$, and hence $\Psi(f_j, f_i) = 0$. Although equation 10 can be approximated using numerical methods, the computational cost can become very expensive; integrating over high-dimensional spaces soon turns intractable even for moderately low values of k. In practice, a solution to this problem is to assume a form of attribute independence as explained next.

B. The Attribute-Independence Approach

Instead of integrating over all attribute space one may look at each attribute independently. In particular, a projection of the data over each attribute transforms the original problem into a new problem made of one-dimensional Gaussian distributions, as shown in Figure 2 (left). We represent the two distributions on attribute A_l, $1 \leq l \leq k$, as $f_j^l(x)$ (corresponding to cluster $\mathcal{K}_j$) and $f_i^l(x)$ (corresponding to class $\mathcal{C}_i$); the parameters for these distributions are easily obtained from the k-dimensional multi-variate Gaussian distributions by extracting the l-entry of the mean vector, and the (l,l)-entry of the covariance matrix.

The computation of the separation of the two one-dimensional distributions, $\Psi(f_j^l, f_i^l) = \Psi_l$, is now performed over a single dimension and is thus less expensive (equation 10). Nevertheless we are now forced to devise a function $\bigcup(\cdot)$ to combine the degree of separation over all attributes:

$$\Psi(f_j, f_i) = \bigcup(\Psi(f_j^1(x), f_i^1(x)), \ldots, \Psi(f_j^k(x), f_i^k(x)) = \bigcup_l \Psi_l \tag{11}$$

Besides the need to define the nature of $\bigcup(\cdot)$, this approach carries a disadvantage. By looking at each attribute independently, two non-overlapping distributions in a k-dimensional space may appear highly overlapped when projected over each attribute, as shown in Figure 2 (right). Our challenge lies on finding an efficient approach to estimate $\Psi(f_j, f_i)$ along a dimension that provides a clear representation of the true separation between objects on both distributions.

C. Projecting Over a Single Dimension Using Fisher Linear Discriminant

Our proposed solution consists of projecting data objects over a single dimension that is orthogonal to Fisher linear discriminant [18]. The general idea is to find a hyperplane that best discriminates data objects in cluster $\mathcal{K}_j$ from data objects in class $\mathcal{C}_i$. The weight vector $\mathbf{w}$ that lies orthogonal to the hyperplane will be used as the dimension upon which the data objects will be projected. The rationale behind this method is that among all possible dimensions over which that data can be projected, classical linear discriminant analysis identifies the vector $\mathbf{w}$ with an orientation that results in a maximum (linear) separation between data objects in $\mathcal{K}_j$ and $\mathcal{C}_i$; the distribution of data objects over $\mathbf{w}$ provide a better indication of the true overlap between $\mathcal{K}_j$ and $\mathcal{C}_i$ in k dimensions compared to the resulting distributions obtained by projecting data objects over the attribute axes. Figure 3 illustrates our methodology. Weight vector $\mathbf{w}$ –which lies orthogonal to the hyperplane that maximizes the separation between the objects in cluster $\mathcal{K}_j$ and class $\mathcal{C}_i$– is used as the dimension over which data objects are projected.

Specifically, Fisher linear discriminant finds the vector $\mathbf{w}$ that maximizes the following criterion function:

$$J(\mathbf{w}) = \frac{\mathbf{w}^t S_B \mathbf{w}}{\mathbf{w}^t S_W \mathbf{w}} \tag{12}$$

The term S_B is also named the between-class scatter matrix; it is simply the outer product of two vectors:

$$S_B = (\mu_j - \mu_i)(\mu_j - \mu_i)^t \tag{13}$$

where μ_j is the mean vector of cluster distribution $f_j(\mathbf{x})$ and μ_i is the mean vector of class distribution $f_i(\mathbf{x})$.

The term S_W is also named the within-class scatter matrix; it is the sum of the scatter matrix over the two distributions:

$$S_W = \sum_{\mathbf{x} \in \mathcal{K}_j} (\mathbf{x} - \mu_j)(\mathbf{x} - \mu_j)^t + \sum_{\mathbf{x} \in \mathcal{C}_i} (\mathbf{x} - \mu_i)(\mathbf{x} - \mu_i)^t \tag{14}$$

Fisher linear discriminant maximizes the ratio of between-class scatter to within-class scatter. Geometrically the goal is to find a vector $\mathbf{w}$ so that the difference of the projected means over $\mathbf{w}$ is large compared to the standard deviations around each mean. It can be shown that a solution maximizing $J(\mathbf{w})$ (equation 12) is in fact independent of S_B:

$$\mathbf{w} = S_W^{-1}(\mu_j - \mu_i) \tag{15}$$

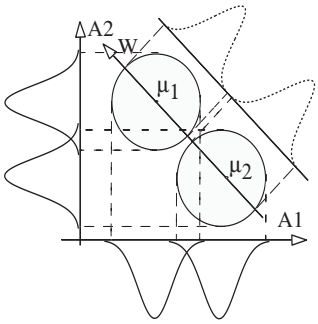

Fig. 3. Weight vector $\mathbf{w}$ –which lies orthogonal to the hyperplane that maximizes the separation between the objects in cluster $\mathcal{K}_j$ and class $\mathcal{C}_i$– is used as the dimension over which data objects are projected.

Data Projection

Projecting data objects over the resulting vector $\mathbf{w}$ obviates working on each attribute separately (as in equation 11); we have then found an efficient approach to estimate the degree of separation between two distributions along a single dimension that captures most of the variability of class $\mathcal{C}_i$ and cluster $\mathcal{K}_j$. To perform the data projection mentioned above we need to transform each original data point $\mathbf{x}$ into its projection x', through a scalar dot product[5], $x' = \mathbf{w}^t\mathbf{x}$.

We will refer to the projected density functions over $\mathbf{w}$ as $f'_i(x)$ (for class $\mathcal{C}_i$) and $f'_j(x)$ (for cluster $\mathcal{K}_j$). Their parameters can be easily estimated after projecting data objects over $\mathbf{w}$. Let μ be the mean of density function $f(\mathbf{x})$, then the projected parameters are defined as

$$\mu' = \mathbf{w}^t\ \mu \qquad \sigma'^2 = \frac{1}{N}\sum(x' - \mu')^2 \tag{16}$$

where μ' and σ'^2 are the projected mean and variance respectively; if the parameters correspond to $f'_j(x)$, N is the number of data objects comprised by cluster $\mathcal{K}_j$ ($N = |\mathcal{K}_j| = N_j$); otherwise N is the number of data objects comprised by class $\mathcal{C}_i$ ($N = |\mathcal{C}_i| = N_i$).

In summary, our approach is to quantify the separation between two one-dimensional Gaussian distributions $f'_j(x)$ and $f'_i(x)$ obtained after projecting data objects in class $\mathcal{C}_i$ and cluster $\mathcal{K}_j$ along vector $\mathbf{w}$.

D. The Distance Between Two One-Dimensional Distributions

To finalize the description of our approach we need only specify how to compute the degree of separation between the two one-dimensional Gaussian distributions f'_j and f'_i, denoted as $\Psi(f'_j, f'_i)$. To that purpose we make use of the concept of relative entropy of two density functions [23]. The relative entropy (or Kullback Leibler distance) between two density functions is the expectation of the logarithm of a likelihood ratio:

$$\Psi(f'_j, f'_i) = D(f'_j||f'_i) = \int f'_j(x)\log\frac{f'_j(x)}{f'_i(x)}dx \tag{17}$$

For our purposes, equation 17 is a measure of the error generated by assuming that function f'_i can be used to represent function f'_j; the integral[6] measures the additional amount of information required to describe cluster $\mathcal{K}_j$ given its most similar class $\mathcal{C}_i$. The higher the distance[7], the higher the amount of information conveyed by cluster $\mathcal{K}_j$.

Since the form of the distributions is known to be Gaussian, we can further simplify our measure (we use natural logarithms to reduce the equation; the resulting information is now expressed in nats instead of bits):

$$\Psi(f'_j, f'_i) = D(f'_j||f'_i) = \int f'_j(x)\ln\frac{f'_j(x)}{f'_i(x)}\ dx \tag{18}$$

$$= \int f'_j \ln \frac{\frac{1}{\sqrt{2\pi\sigma'^2_j}}\ e^{-\frac{(x-u'_j)^2}{2\sigma'^2_j}}}{\frac{1}{\sqrt{2\pi\sigma'^2_i}}\ e^{-\frac{(x-u'_j)^2}{2\sigma'^2_i}}}\ dx \tag{19}$$

$$= \int f'_j \ln(\frac{\sigma'_i}{\sigma'_j} e^{\frac{1}{2}(\frac{(x-u'_j)^2}{\sigma'^2_i} - \frac{(x-u'_j)^2}{\sigma'^2_j})})\ dx \tag{20}$$

$$= \int f'_j \ln \frac{\sigma'_i}{\sigma'_j}\ dx + \frac{1}{2\sigma'^2_i}\int f'_j(x-u'_i)^2\ dx - \frac{1}{2\sigma'^2_j}\int f'_j(x-u'_j)^2\ dx \tag{21}$$

$$= \ln \frac{\sigma'_i}{\sigma'_j}\int f'_j\ dx + \frac{1}{2\sigma'^2_i}\int f'_j(x-u'_i)^2\ dx - \frac{\sigma'^2_j}{2\sigma'^2_j} \tag{22}$$

$$= \ln \frac{\sigma'_i}{\sigma'_j} + \frac{1}{2}[\frac{1}{\sigma'^2_i}\int (x-u'_i)^2 f'_j\ dx - 1] \tag{23}$$

$$= \ln \frac{\sigma'_i}{\sigma'_j} + \frac{1}{2}[\frac{\mathtt{E}_{f'_j}[(x-u'_i)^2]}{\sigma'^2_i} - 1] \tag{24}$$

Equation 24 shows the behavior of relative entropy over two one-dimensional Gaussian distributions. If both distributions are the same, then the expectation of $(x-u'_i)^2$ according to distribution f'_j is identical to σ'^2_i and $\Psi(f'_j, f'_i) = 0$. As the two distributions differ the value of $\Psi(f'_j, f'_i)$ grows above zero.

E. Overview of our Approach and Computational Complexity

To summarize, our approach is divided into two steps:

1) Projecting data objects in cluster $\mathcal{K}_j$ and class $\mathcal{C}_i$ over the weight vector $\mathbf{w}$ that lies orthogonal to the hyperplane that maximizes the separation between $\mathcal{K}_j$ and $\mathcal{C}_i$ (section III-C), and
2) Computing the degree of separation between the resulting one-dimensional Gaussian distributions (section III-D).

Figure 4 provides an algorithmic description of our method. The computational complexity of the algorithm is dominated by the first step (Fig. 4, lines 1-7) where the goal is to find the weight vector $\mathbf{w}$. The most expensive calculation is that of the within-class scatter matrix S_W and its inverse. The complexity is of order $O(k[N_j + N_i]^2)$, where k is the number of attributes and $N_j + N_i$ is the total number of data objects comprised by cluster $\mathcal{K}_j$ and class $\mathcal{C}_i$. Even though the computational cost is quadratic on $N_j + N_i$, we expect the number of data objects on both cluster and class to be mush less than the total size of the data set (i.e., we expect $N_j + N_i \ll N$).

On a pentium 4 processor with 1GB of memory, the execution time for the two steps mentioned above on a dataset corresponding to Martian landscapes with 386 data objects (section IV) is on average less than one second.

F. Preliminary Assessment

We report on a preliminary assessment using artificial datasets comparing our method with the attribute-independence approach (section III-B). Our artificial datasets comprise data objects (i.e. points) drawn from two Gaussian distributions with different means but same standard deviation on a two-dimensional attribute space. The location of the means is selected as follows. One mean is chosen uniformly randomly on the plane, while the other mean is randomly located away from the first mean at a fixed distance (e.g., one standard deviation). Our experiments vary the number of points drawn from each distribution and the distance between the means.

The degree of separation under the attribute independence approach simply averages the relative entropy (or Kullback Leibler distance) of the distributions obtained after projecting the data on each attribute. The degree of separation is then as follows (equation 11):

Algorithm 1: External Cluster Assessment
Input: cluster $\mathcal{K}_j$, class $\mathcal{C}_i$
Output: Distance $\Psi(f'_j, f'_i)$
DISTANCE($\mathcal{K}_j$,$\mathcal{C}_i$)
(1) Estimate mean vector μ_j (cluster $\mathcal{K}_j$)
(2) Estimate mean vector μ_i (class $\mathcal{C}_i$)
(3) Compute the within-class scatter matrix:
(4) $S_W = \sum_{\mathbf{x}\in\mathcal{K}_j}(\mathbf{x}-\mu_j)(\mathbf{x}-\mu_j)^t + \sum_{\mathbf{x}\in\mathcal{C}_i}(\mathbf{x}-\mu_i)(\mathbf{x}-\mu_i)^t$
(5) Find the vector $\mathbf{w} = S_W^{-1}(\mu_j - \mu_i)$
(6) Find projected densities over $\mathbf{w}$:
(7) $f'_i(x)$ (class $\mathcal{C}_i$) and $f'_j(x)$ (cluster $\mathcal{K}_j$)
(8) Compute $\Psi(f'_j, f'_i) = D(f'_j||f'_i) = \int f'_j(x) \ln \frac{f'_j(x)}{f'_i(x)}\, dx$
(9) **return** $\Psi(f'_j, f'_i)$

Fig. 4. The logic behind our approach to external cluster assessment.

$$\Psi(f_i, f_j) = \bigcup(\Psi(f_i^1(x), f_j^1(x)), \ldots, \Psi(f_i^k(x), f_j^k(x)) = \frac{1}{k} D(f'_j||f'_i) \tag{25}$$

where f'_j and f'_i are obtained by projecting the data objects on each attribute.

Figure 5 shows our results. On all graphs, the x-axis stands for the size of the dataset (on a logarithmic scale); the y-axis stands for the degree of separation (i.e., relative entropy) between both Gaussian distributions. Each result is the average of ten runs; we show 95% confidence intervals (using a t-student distribution); the solid line corresponds to the true degree of separation assuming an infinite sample.

Our method takes slightly longer to converge when the distance between the means is zero (i.e., when both cluster and class belong to the same distribution). This is the result of finding a vector $\mathbf{w}$ orthogonal to Fisher's linear discriminant when no decision boundary actually exists (Fig. 5 top-left). As the distance between the means grows larger, however, our method converges to the true separation relatively fast. In contrast, the attribute-independence approach tends to underestimate the true degree of separation; attribute projections show a distorted view of the true overlap between the two distributions over the plane. In summary, our method outperforms the attribute-independence approach when projections over the attribute axes convey a distorted view of the actual location of the class-cluster pair.

IV. CHARACTERIZATION OF MARTIAN SURFACES

We now turn to an area of application where our approach is tested. Our study revolves around the characterization and classification of Martian surfaces. We study a large set of Martian sites showing various types of surfaces. First, we discuss the notion of geological units - the standard classification of Martian sites assigned by domain experts (geologists) after careful examination of a site's image. Assigning geological units to each site in our set divides the dataset into m predefined external classes. Second, we discuss the notion of a network descriptor - a numerical attribute of a Martian site that is calculated from its topography. Network descriptors are 4-dimensional vectors. Applying a clustering algorithm to network descriptors partitions the dataset into n clusters. Our metric is used to assess the distance between those clusters and the set of classes predefined on the basis of geological units.

A. External Classes: Geological Units

Presently, the main tool for studying the Martian surface (and other planetary surfaces) is geologic mapping [15]. A geologic map is a 2-dimensional projection of the 3-dimensional distribution of geological units, bodies of rock that are thought to have formed by a particular process or set of related processes over a discrete time span [24]. In a terrestrial context, geological units are determined from in situ inspections. In a Martian context, however, these units are determined from images through topographical expressions. A geologic map is a thematic

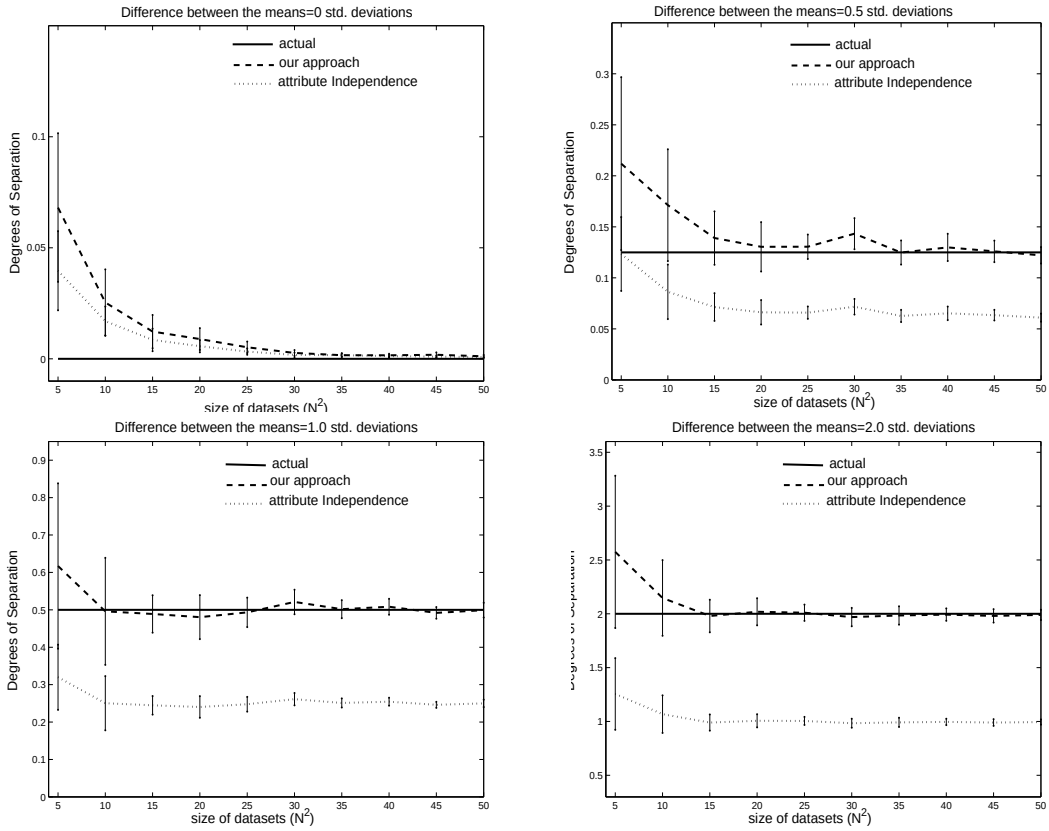

Fig. 5. A comparison of our approach with the attribute independence approach on four artificial datasets. The difference between the means varies from zero standard deviations (upper left) up to two standard deviations (bottom right).

map of geological units, an encapsulation of a huge amount of information into a concise output by means of human interpretation.

Figure 6 shows an example of a geologic map. The East Mangala Valles region on Mars (coordinates of the site's center are: -147.56E, 9.95S) has been manually mapped (middle panel) from imagery data (left panel). The site has been divided [25] into geological units indicated by different colors and labeled in the legend (right panel). The criteria considered by a geologist in making this division includes terrain texture, geological structure, age, and stratigraphy. The labels given to geological units are shortcuts for longer natural language descriptions. For example, the unit Npl1 is described as "highly cratered uneven surface of moderate relief; fractures, faults, and small channels common." The vast majority of geological units have names that start with letters N, H, or A indicating Noachian, Hesperian, and Amazonian stratigraphic epochs, respectively. However, sometimes mappers encounter a terrain that is specific to a given site and assign it a name outside of the general framework. An example of such assignment are units C1 to C4 on Figure 6.

B. Quantitative Characterization of Martian Surfaces

Geological units are the traditional, qualitative means of classifying Martian surfaces. One shortcoming of such classification is that it cannot be automated. Given the vast amount of data collected by spacecrafts, the field of Martian geomorphology would benefit from an automated, quantitative classification of surfaces. A stumbling block to the development of such an automated classification is the lack of an adequate yet concise mathematical representation of a topographic surface. It has been proposed [17] that a binary tree data structure (tantamount to a terrain's "drainage" network) provides such a representation. We explain such representation next.

An automated classification of Martian surfaces uses digital topography data. Martian topography data was gathered by the Mars Orbiter Laser Altimeter (MOLA) instrument [26]. This data was subsequently used to construct the Mission Experiment Gridded Data Records (MEGDR) [27] which are global topographic maps of Mars with

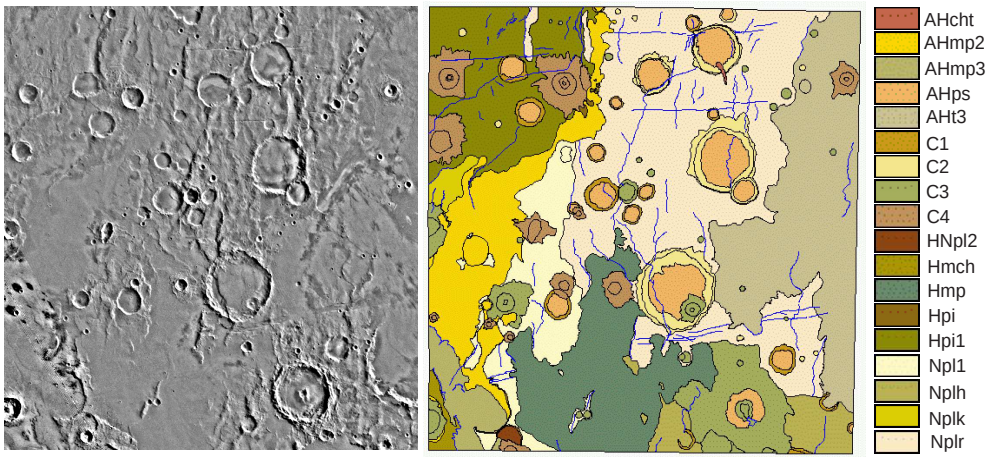

Fig. 6. (left) Image of East Mangala Valles region on Mars. The width of this side is ≈ 340 km. (middle) The geologic map of Mangala Valles region. Different geological units are indicated by different colors. (right) The legends for the geologic map on the left.

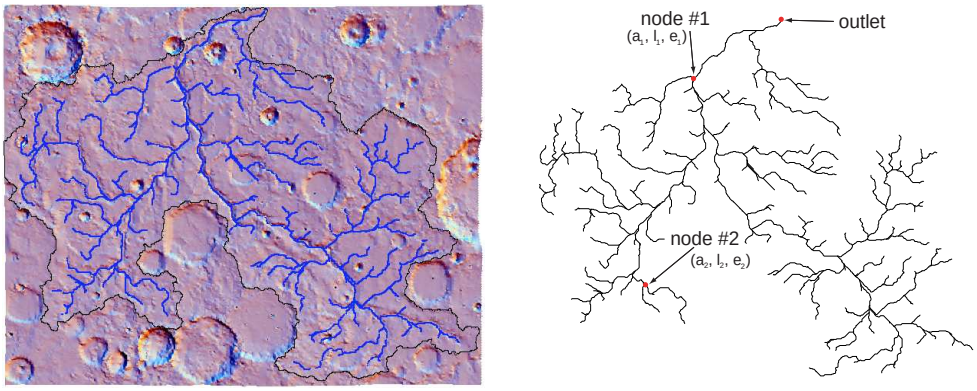

Fig. 7. (left) Visual rendering of an elevation field of Naktong Vallis region on Mars (31.3E, 6.6N). The black line shows the boundary of the catchment and the blue lines show the drainage network of arbitrary penetration into the catchment. (right) The binary tree representing the "drainage" network. Red dots indicate three points of interest, an outlet and two out of 145 nodes. Values of a, l, and e are calculated and stored at the nodes of the tree.

a resolution of ≈ 0.5 km. For a site of interest the MEGDR is used to construct a digital elevation model (DEM) of the site. The DEM is a regular grid of cells with assigned elevation values. A hydrology-inspired algorithm was designed [17] for quantitative analysis of surfaces as represented by DEMs. The algorithm can be thought of as subjecting a surface to "artificial rain" and registering how it drains. The term "drain" is used here as a metaphor for connectivity between different points on the surface. The resultant drainage pattern characterizes the texture and structure of the surface.

Specifically, a point called an outlet is selected and the portion of the surface that ultimately drains through this point is called a catchment. A drainage network is the part of the catchment where the flow is concentrated. The extent of penetration of the network into the catchment is adjustable, the network can reach into every cell in the catchment. The network has a spanning binary tree geometry with an outlet being at the root of the tree. Figure 7 illustrates a relation between the surface, the catchment, and the drainage network.

The binary tree network doubles as a data structure with every node S holding values of three variables: a, an area of catchment with an outlet at S; l, length of the longest upstream path starting at S; e, potential energy dissipated along a segment of the network terminating at S. We describe the network, and thus the catchment, and ultimately the surface in terms of probability distribution functions of these three variables. Reflecting the fractal structure of the network, all three variables have power law distributions, $P(a) \propto a^{-(1+\tau)}$, $P(l) \propto l^{-(1+\gamma)}$, $P(e) \propto e^{-(1+\beta)}$, and a network can be statistically characterized by the power law indices τ, γ, and β. An additional variable, ρ, the uniformity of drainage density [16] is added to the three power law indices to form a 4-dimensional

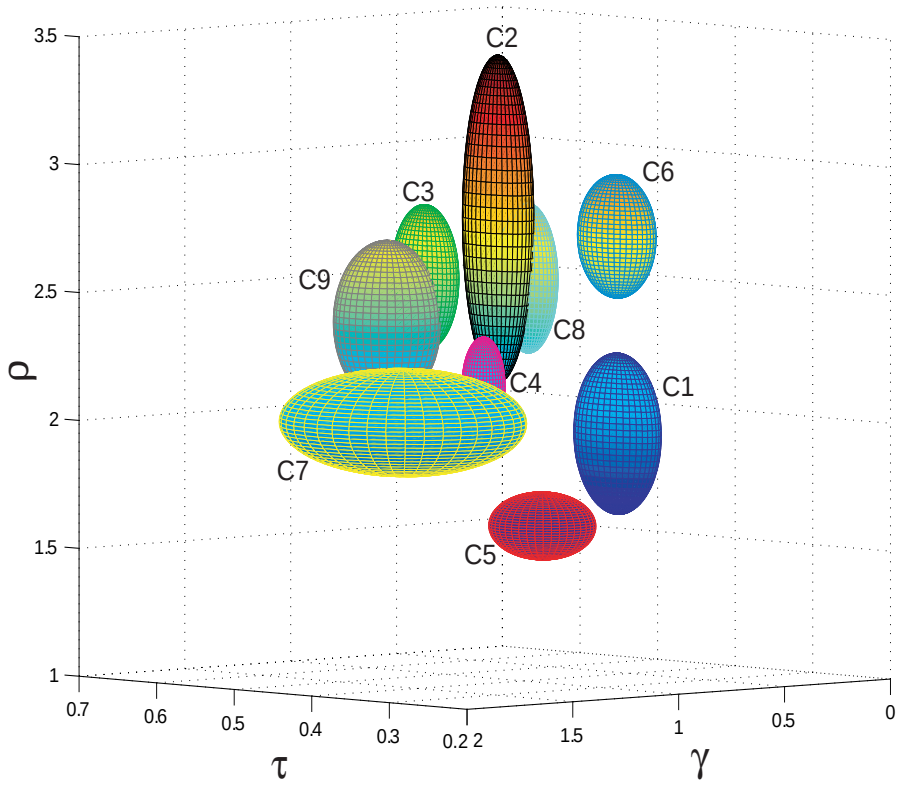

Fig. 8. Nine clusters resulting from partition of dataset of Martian sites with respect to the values of their network descriptors. Ellipsoids represent 3-dimensional projections of clusters in the 4-dimensional space.

vector $(\tau, \gamma, \beta, \rho)$ which we call a network descriptor. A network descriptor provides an algorithmically derived, quantitative characterization of a surface that is independent from a descriptive characterization using geological units.

V. Empirical Study

Our dataset consists of 386 Martian sites taken from a wide range of Martian latitudes and elevations. They represent all three major epochs and are classified into $m = 16$ different geological units (classes): Npl1(28), Npl2(17), Npld(41), Nple(8), Nplr(31), Nh1(11), Had(15), Hh3(12), HNu(16), Hpl3(14), Hr(72), Hvk(32), Ael1(10), Aoa(15), Apk(38), and Aps(26). The numerical values between parentheses indicate the number of sites in a given class. We have clustered the dataset of 386 Martian sites with respect to the similarity of their network descriptors. Our empirical study is divided into three steps: 1) an internal assessment of the quality of the clusters alone; 2) an external cluster assessment by looking at the separation between clusters and classes (geological units) using our proposed approach; and 3) a geomorphic interpretation of the clusters.

A. Assessing the Quality of Clusters Alone

We cluster the dataset of Martian sites with respect to their network descriptors using a probabilistic clustering algorithm. The algorithm assumes a data object $\mathbf{x}$ belongs to a cluster $\mathcal{K}_j$ with a posterior probability $P(\mathcal{K}_j|\mathbf{x})$. Object $\mathbf{x}$ is assigned to the cluster exhibiting highest posterior probability (i.e., object $\mathbf{x}$ belongs to cluster $\mathcal{K}_j$ if $P(\mathcal{K}_j|\mathbf{x}) > P(\mathcal{K}_l|\mathbf{x})$, $l = 1 \ldots n$, $l \neq j$).

The algorithm works under a Bayesian framework. The posterior probability of a cluster $\mathcal{K}_j$ given an example $\mathbf{x}$ is expressed as follows:

$$P(\mathcal{K}_j|\mathbf{x}) = \frac{P(\mathbf{x}|\mathcal{K}_j)P(\mathcal{K}_j)}{\sum_{l=1}^{n} P(\mathbf{x}|\mathcal{K}_l)P(\mathcal{K}_l)} \tag{26}$$

Since the denominator is constant for all clusters we can dispense with it. We then assign cluster $\mathcal{K}_j$ to example $\mathbf{x}$ if it maximizes $P(\mathbf{x}|\mathcal{K}_j)P(\mathcal{K}_j)$. This requires an estimation of the parameter vector θ_j characterizing $P(\mathbf{x}|\mathcal{K}_j)$ (if assuming a Gaussian distribution $\theta_j = [\mu_j, \Sigma_j]$), and the a priori probability $P(\mathcal{K}_j)$. Such estimations can be performed using the Expectation Maximization (EM) technique [28]. Since the number n of clusters is assumed to be known, the algorithm tries different values using cross-validation[8].

The dataset of 386 Martian sites was partitioned into nine clusters labeled C1(23), C2(28), C3(37), C4(49), C5(29), C6(35), C7(16), C8(129), and C9(40). The numerical values between parentheses indicate the number of sites in a given cluster. Each cluster can be represented as a 4-dimensional ellipsoid in the $(\gamma, \tau, \beta, \rho)$ space. The center of an ellipsoid is at $(\langle\gamma\rangle, \langle\tau\rangle, \langle\beta\rangle, \langle\rho\rangle)$, where the means are calculated over the objects belonging to a given cluster. For visualization purposes, the length of each ellipsoid's semi-axis is equivalent to one standard deviation (extracted from the diagonal of the covariance matrix).

Figure 8 shows a projection of ellipsoids representing all nine clusters onto the (γ, τ, ρ) space[9]. The clusters are well separated in the space indicating that our dataset has been divided into distinct groups. Similar projections onto the three other possible 3-dimensional sub-spaces confirm this conclusion. To assess quantitatively the quality of our clusters we have calculated a 9×9 matrix of Kullback-Leibler distances between the clusters (following the methodology of section III). Of course, the diagonal entries in this matrix are all equal to zero. The smallest off diagonal entry, corresponding to the distance between the clusters[10] C3 and C9 equals 1.45. Even this smallest distance indicates a significant separation (see Table 2). The largest off diagonal entry, corresponding to the distance between the clusters C7 and C8, equals 52.84. The average distance is 11.18, and the standard deviation is 9.65. Thus, our clustering of Martian sites resulted in a meaningful classification. A physical interpretation of this classification is attempted in section 5.3.

B. Comparing Clusters to Geological Units

We now assess the degree of separation between the nine clusters and the sixteen Martian geological units (classes). The network descriptors for the sites classified into a single geological unit form a "concentration" in the $(\gamma, \tau, \beta, \rho)$ space. Such a concentration can be represented as a 4-dimensional ellipsoid employing the method used in section 5.1 for cluster representation. Conceptually, the comparison between the clusters and the classes amounts to assessing the degree of overlapping between the sets of corresponding ellipsoids. In practice, the assessment is achieved using our proposed approach (section III-D). We have calculated a 9×16 matrix of Kullback-Leibler distances between the clusters and classes. The distances vary from a minimum of 0.3078 (between C4 and Hr) to a maximum of 17.97 (between C7 and Had). The average distance is 2.97 and the standard deviation is 2.67. Table 1 shows the Kullback-Leibler distances between clusters and selected classes. The first column corresponds to the nine clusters obtained by partitioning the dataset of Martian sites on the basis of similarity between network descriptors. For each row, the second column corresponds to the class with smallest separation to that cluster, the third column corresponds to the class with the second smallest separation, and so on. We report on the five classes with smallest separation for each cluster. Within parentheses we show the identification label (the name of the geological unit) for each class. As a baseline for comparison, Table 2 shows the degree of separation using our proposed approach between two one-dimensional Gaussian distributions having the same variance. The columns indicate the difference between the means in units of standard deviation.

The results in Table 1 indicate that none of the nine clusters can serve as a surrogate for any geological units. For a cluster to be consider a candidate for class surrogate, its Kullback-Leibler distance to that class should be small, and its distances to all other classes should be large. Since none of the nine clusters meets such criteria, we conclude our results point to a new classification of Martian sites. A deeper analysis of Table 1 shows that cluster C4 has a relatively small separation values from a number of classes: Hr, Npl1, Nplr, and Npl2. These separation values have similar magnitudes, but none stands out as significantly smaller than the others. Closer examination reveals that sites in those four different classes are distributed similarly in the $(\gamma, \tau, \beta, \rho)$ space. The Kullback-Leibler distances between pairs of these classes are all smaller than 0.34. Thus, the ellipsoids representing Hr, Npl1, Nplr, and Npl2 are all very similar to each other. The smaller ellipsoid representing C4 is located inside the other four ellipsoids. This geometry explains why the separation between cluster C4 and the other four classes is similar and small. Clearly, cluster C4 groups catchments that occur often in Martian terrain classified as Hr, Npl1, Nplr, and Npl2. However, the differences between these surfaces, previously identified by geologists, are not readily encapsulated

TABLE I

A MEASURE OF THE DEGREE OF SEPARATION BETWEEN CLUSTERS AND CLASSES IN THE CONTEXT OF MARTIAN TOPOGRAPHY.

	Geological Units				
Clusters	Most Similar	2nd	3rd	4rd	5th
C1	1.4622 (Hr)	1.489 (Nplr)	1.502 (Npld)	1.5983 (Hh3)	1.8741 (Npl1)
C2	0.6015 (Aoa)	0.7812 (Hpl3)	0.823 (Nple)	0.8438 (Hvk)	1.4669 (Ael1)
C3	0.9738 (Nh1)	0.9908 (Apk)	1.0827 (Npl1)	1.1275 (Aps)	1.1356 (Hvk)
C4	0.3078 (Hr)	0.3584 (Npl1)	0.4127 (Nplr)	0.5756 (Npl2)	0.7287 (Aps)
C5	0.8789 (Hh3)	1.6257 (Nplr)	1.6997 (Npl1)	2.011 (Hr)	2.0254 (Npld)
C6	1.0818 (Hvk)	1.3423 (Ael1)	1.7171 (Hpl3)	1.7177 (Had)	1.8053 (Npld)
C7	1.1037 (Nplr)	1.7299 (Npl1)	3.1915 (Nple)	3.6975 (Npl2)	5.038 (HNu)
C8	0.3461 (Hvk)	0.4909 (Ael1)	0.5942 (Apk)	0.7608 (Aps)	0.764 (HNu)
C9	0.9535 (Nplr)	1.2416 (Nh1)	1.2812 (Aps)	1.3445 (Hpl3)	1.3565 (Hr)

TABLE II

A MEASURE OF THE DEGREE OF SEPARATION BETWEEN TWO ONE-DIMENSIONAL GAUSSIAN DISTRIBUTIONS, f_1 AND f_2, WITH EQUAL VARIANCE.

	Difference between the means										
	0.2 σ	0.4 σ	0.6 σ	0.8 σ	1.0 σ	1.2 σ	1.4 σ	1.6 σ	1.8 σ	2.0 σ	2.2 σ
$\Psi(f_1, f_2)$	0.02	0.08	0.18	0.32	0.50	0.72	0.98	1.28	1.62	2.00	2.42

by network descriptors. The most populous cluster C8 groups catchments that are typical for many surfaces. This is why it also has relatively small separations from a number of classes. Its average distance from all 16 classes is 0.97 with a standard deviation of 0.41. On the other hand, cluster C1 groups peculiar catchments that are not common on any surfaces. These are interiors of large craters. As a result, cluster C1 is well separated from all classes. Its average distance from all 16 classes is 3.22 with a standard deviation of 1.51.

C. Physical Interpretation of Clusters

Using our method for external cluster assessment, we were able to determine that partitioning the dataset of Martian sites on the basis of network descriptors produced a novel classification that does not match the traditional classification based on geological units. In general, the new classification pertains to the character of catchments. The most populous cluster, C8, groups sites with network descriptors describing a catchment that has a character common to many Martian (and terrestrial) terrains. This character could be succinctly described as moderate elongation. In contrast, cluster C9 groups sites with network descriptors indicating “square” catchments without much elongation; cluster C6 groups sites with narrow, elongated catchments. It remains an open question to explain how the shape of a catchment relates to terrain attributes such as texture, structure, and stratigraphy.

Figure 9 shows an example of the difference between catchment shapes and more traditional geomorphic attributes. Four Martian surfaces are shown in a 2×2 matrix arrangement. Surfaces in the same row belong to the same geological unit, surfaces in the same column belong to the same cluster. The top two sites show two surfaces from the Hr geological unit that is described as ”ridged plains, moderately cratered, marked by long ridges.” These features can indeed be seen in the two surfaces. Despite such texture similarity they have very different catchments as indicated by their drainage networks. The bottom two sites show two surfaces from the Apk unit described as ”smooth plain with conical hills or knobs.” Again, looking at Figure 9 it is easy to see the similarity between these two surfaces based on that description. Nevertheless, the two terrains have catchments with markedly different character. On the basis of catchment similarity, these four surfaces could be divided vertically instead of horizontally. Such division corresponds to our cluster partition.

VI. SUMMARY AND CONCLUSIONS

Clustering algorithms arrange data objects into groups that convey potentially meaningful and novel domain interpretations. When the same data objects have been previously framed into a particular classification scheme, the value of each cluster can be assessed by estimating the degree of separation between the cluster and its most similar class. In this paper we propose an approach to external cluster assessment based on modeling each cluster and class as a multivariate Gaussian distribution; the degree of separation between both distributions follows an

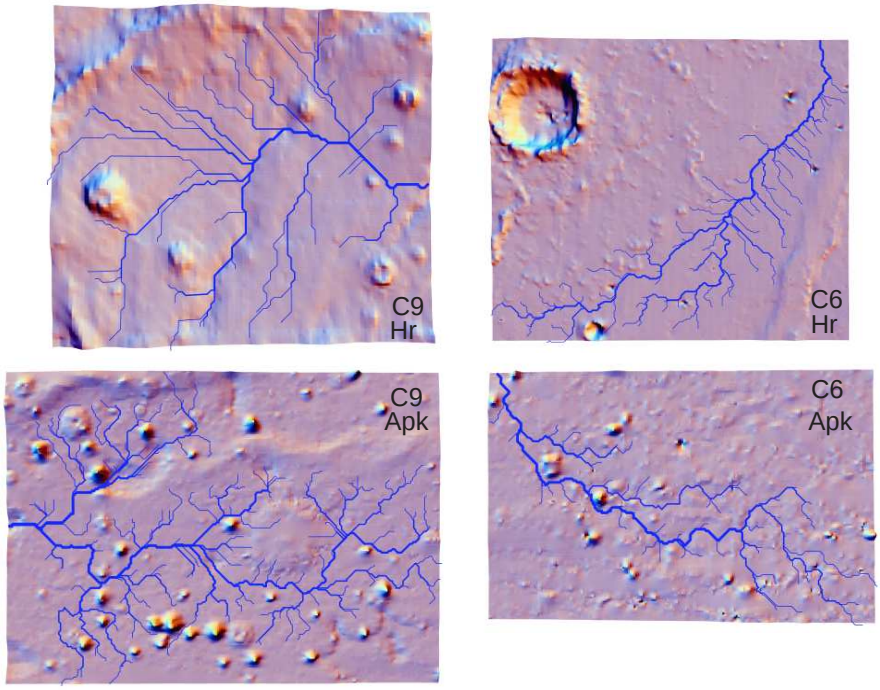

Fig. 9. Four Martian surfaces from two different geological units and belonging to two different clusters. Binary trees representing "drainage" networks are drawn on top of the surfaces.

information-theoretic measure known as relative entropy or Kullback Leibler distance. Compared to previous work, our method evaluates each cluster individually and employs a probabilistic model (as opposed to a contingency table) in estimating the separation between class and cluster.

Our approach achieves a balance between the computational cost of approximating the separation of two distributions when integrating over the whole attribute space, as compared to integrating over each attribute independently. In the first case, the cost of integrating over high dimensional spaces soon turns intractable even for moderately low number of attributes. In the second case, two non-overlapping distributions in the attribute space may appear highly overlapped when projected over each attribute. Our approach estimates the separation of two distributions along a single dimension, by projecting all data objects over the vector that lies orthogonal to the hyperplane that maximizes the separation between cluster and class (using Fisher's Linear Discriminant).

We test our approach on a dataset of Martian surfaces by comparing their description-based classification into geological units with a new, algorithm-based division. Using our approach we have determined that a particular automated classification, based on hydrology-inspired variables, cannot be used in place of geological units. Instead, we discovered the Martian dataset can be divided into high quality clusters with respect to these novel variables.

Future work will assess the value of clusters obtained with alternative algorithms (other than the probabilistic algorithm used in section V-A). We also plan to devise better modelling techniques for the external class distribution. Our approach assumes each cluster can be modelled through a multivariate Gaussian distribution; while this assumption is reasonable due to the expected local nature of each cluster, the same assumption comes unwarranted for external classes (their nature is often unknown). An alternative approach is to model each external class as a mixture of models. Finally, one line of research is to design clustering algorithms that search for clusters in a direction that maximizes a metric of relevance or *interestingness* as dictated by an external classification scheme. Specifically, a clustering algorithm can be set to optimize a metric that rewards clusters exhibiting little (conversely strong) resemblance to existing classes.

Acknowledgments

Thanks to the Lunar and Planetary Institute, which is operated by USRA under contract CAN-NCC5-679 with NASA, for facilitating data on Martian landscapes. The paper is LPI contribution No. 1237. This material is

based upon work supported by the National Science Foundation under Grants no. IIS-0431130, IIS-0448542, and IIS-0430208.

Footnotes

1) Department of Computer Science, University of Houston. 4800 Calhoun Rd., Houston TX 77204-3010, USA.
2) Lunar and Planetary Institute. 3600 Bay Area Blvd, Houston TX 77058-1113, USA.
3) A third type of cluster validation, called *relative*, compares different clustering structures obtained from the same clustering algorithm [8].
4) We consider a flat type of clustering (as opposed to hierarchical) where each object is assigned to exactly only cluster.
5) The projections have a clear geometrical interpretation when performed over $\mathbf{w}_0 = \frac{\mathbf{w}}{||\mathbf{w}||}$. Vector $\mathbf{w}_0$ is a normalized vector (i.e., $||\mathbf{w}_0|| = 1$). But the magnitude of $\mathbf{w}$ is of no consequence; if $||\mathbf{w}|| \neq 1$ the result is simply a change on the scale of x'.
6) The integral runs along values of x for which $f'_j(x) > 0$ (i.e., runs along the support set of f'_j). We assume that $0 \log \frac{0}{0} = 0$, and that $\forall x(f'_j(x) > 0) \rightarrow (f'_i(x) > 0)$ (i.e., the support set of f'_j is embedded in the support set of f'_i).
7) Note that although $D(f'_j||f'_i) \geq 0$,
relative entropy is not a true distance because it is not symmetric [23]; that is $D(f'_j||f'_i) \neq D(f'_i||f'_j)$.
8) The algorithm is part of the WEKA machine-learning tool [29].
9) We use a projection to facilitate visualization of our clusters.
10) The matrix is not symmetric as $D(f'_j||f'_i) \neq D(f'_i||f'_j)$. When referring to the distance between two clusters CA and CB we assume that particular order (i.e., we assume $D(f'_A||f'_B)$).

REFERENCES

[1] P. Diggle, *Statistical Analysis of Spatial Point Patterns*. Academic Press, 1983.
[2] A. Jain and R. Dubes, *Algorithms for Clustering Data*. Prentice Hall, 1988.
[3] e. Panayirci and R. Dubes, "A test for multidimensional clustering tendency," *Pattern Recognition*, vol. 16, no. 4, pp. 433–444, 1983.
[4] B. Ripley, *Spatial Statistics*. John Wiley & Sons, 1981.
[5] G. Zeng and R. Dubes, "A comparison of tests for randomness," *Pattern Recognition*, vol. 18, no. 2, pp. 191–198, 1985.
[6] W. M. Rand, "Objective criterion for evaluation of clustering methods," *Journal of American Statistical Association*, vol. 66, pp. 846–851, 1971.
[7] E. Fowlkes and C. Mallows, "A method for comparing two hierarchical clusterings," *Journal of American Statistical Association*, vol. 78, pp. 553–569, 1983.
[8] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*. Academic Press, 2003.
[9] R. Krishnapuran, H. Frigui, and O. Nasraoui, "Fussy and possibilistic shell clustering algorithms and their application to boundary detection and surface approximation, part ii," *IEEE Transactions on Fuzzy Systems*, vol. 3, no. 1, pp. 44–60, 1995.
[10] F. Rolph and D. Fisher, "Test for hierarchical structure in random data sets," *Systematic Zoology*, vol. 17, pp. 407–412, 1968.
[11] B. Dom, "An information-theoretic external cluster-validity measure," *Research Report, IBM T.J. Watson Research Center RJ 10219*, 2001.
[12] T. Kanungo, B. Dom, W. Niblack, and D. Steele, "A fast algorithm for mdl-based multi-band image segmentation," in *Image Technology*, J. Sanz, Ed. Springer-Verlag, 1996.
[13] G. W. Milligan, S. C. Soon, and L. M. Sokol, "The effect of cluster size, dimensionality, and the number of clusters on recovery of true cluster structure," *IEEE Transactions on Patterns Analysis and Machine Intelligence*, vol. 5, no. 1, pp. 40–47, 1983.
[14] P. Cheeseman and J. Stutz, "Bayesian classification (autoclass): Theory and results," in *Advances in Knowledge Discovery and Data Mining*, U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Eds. AAAI Press/MIT Press, 1996.
[15] D. E. Wilhelms, "Planetary mapping," R. Greeley and R. Batson, Eds. Cambridge, UK, Cambridge Univ. Press., 1990.
[16] T. Stepinski, M. M. Marinova, P. McGovern, and S. M. Clifford, "Fractal analysis of drainage basins on mars," *Geophysical Research Letters*, vol. 29, no. 8, 2002.
[17] T. e. a. Stepinski, "Martian geomorphology from fractal analysis of drainage networks," *Journal of Geophysical Research*, vol. 109, no. E02005, 10.1029/2003JE0020988, 2004.
[18] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. Wiley Ed. 2nd Edition, 2001.
[19] L. Hubert and J. Schultz, "Quadratic assignment as a general data analysis strategy," *British Journal of Mathematical and Statistical Psychology*, vol. 29, pp. 190–241, 1976.
[20] S. Vaithyanathan and B. Dom, "Model selection in unsupervised learning with applications to document clustering," in *Proceedings of the Sixteenth International Conference on Machine Learning*, Stanford University, CA., 2000.
[21] Y. Shreider, *Method of Statistical Testing: Monte Carlo Method*. Elsevier North-Holland, 1964.
[22] I. Sobo, *The Monte Carlo Method*. Mir Publishers, 1984.
[23] T. M. Cover and J. Thomas, *Elements of Information Theory*. Wiley-Interscience, 1991.

[24] K. Tanaka, “The venus geologic mappers’ handbook,” *U.S. Geol. Surv. Open File Rep.*, pp. 99–438, 1994.
[25] M. G. Chapman, H. Masursky, and A. L. J. Dial, “Geological map of science area 1a, east mangala valles region on mars,” *U.S.G.S. Misc. Geol. Inv., Map I-1696*, 1989.
[26] M. Zuber, D. Smith, S. Solomon, D. Muhleman, J. Head, J. Garvin, J. Abshire, and J. Bufton, “The mars observer laser altimeter investigation,” *J. Geophys. Res.*, vol. 97, pp. 7781–7797, 1992.
[27] D. Smith, G. Neumann, R. Arvidson, E. Guinness, and S. Slavney, “Global surveyor laser altimeter mission experiment gridded data record,” *NASA Planetary Data System, MGS-M-MOLA-5-MEGDR-L3-V1.0.*, 2003.
[28] G. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*. John Wiley and Sons, 1997.
[29] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Academic Press, 2000.