

An Quantification of Cluster Novelty with an Application to Martian Topography

R. Vilalta¹, T. Stepinski², M. Achari¹, and F. Ocegueda-Hernandez³

¹ Department of Computer Science, University of Houston
4800 Calhoun Rd., Houston TX 77204-3010, USA
{vilalta, amkchari}@cs.uh.edu

² Lunar and Planetary Institute
3600 Bay Area Blvd, Houston TX 77058-1113, USA
tom@lpi.usra.edu

³ CINVESTAV
López Mateos Sur 590, Guadalajara, Jalisco, C.P. 45090, México
focegued@gdl.cinvestav.mx

Abstract. Automated tools for knowledge discovery are frequently invoked in databases where objects already group into some known classification scheme. In the context of unsupervised learning or clustering, such tools delve inside large databases looking for alternative classification schemes that are both meaningful and novel. A quantification of cluster novelty can be looked upon as the degree of separation between each new cluster and its most similar class. Our approach models each cluster and class as a Gaussian distribution and estimates the degree of overlap between both distributions by measuring their intersecting area. Unlike other metrics, our method quantifies the novelty of each cluster individually, and enables us to rank classes according to its similarity to each new cluster. We test our algorithm on Martian landscapes using a set of known classes called geological units; experimental results show a new interpretation for the characterization of Martian landscapes.

1 Introduction

Clustering algorithms are useful tools in revealing structure from unlabelled 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 (e.g., sum-of-squared errors, minimum variance), and the similarity measure (e.g, Euclidean distance). In real-world applications, however, the discovery of natural groups of data objects is often of limited use; in addition one needs to assess the quality of the resulting clusters against known classifications. An understanding of the output of the clustering algorithm can be achieved by either finding a resemblance of the clusters with existing classes, or if no resemblance is found, by providing an interpretation to the new groups of data objects.

This paper proposes a method to assess the novelty of a set of clusters under the assumption of the existence of a known classification of objects. Most previous 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; approaches vary in nature from information-theoretic [1, 4] to statistical [7, 3, 6]. By averaging the degree of match across all classes and clusters, such metrics fail to identify the potential novelty of single clusters. Moreover, the lack of a probabilistic model in the representation of data distributions precludes inferring the extent to which a class-cluster pair intersect. Our goal is to be able to identify the existence of novel clusters by looking at each of them individually, ranking all classes against each cluster based on their degree of overlap or intersection.

We test our methodology on a database containing images of Mars landscapes produced by the Mars Orbiter Laser Altimeter (MOLA) [9]. Each terrain is characterized through a computational analysis of its drainage networks and represented as a real vector. We apply a probabilistic clustering algorithm that groups terrains into clusters by modelling each cluster through a probability density function; each terrain (i.e., each vector) in the database has a probability of class membership and is assigned to the cluster with highest posterior probability (Section 4). We assess the novelty of the output clusters by applying our proposed methodology using a known classification of Mars surface based on regions known as geological units. The analysis has prompted a new classification of Mars landscapes based on hydrological aspects of landscape morphology.

This paper is organized as follows. Section 2 provides background information and defines current metrics that compare sets of clusters with known object classifications. Section 3 explains our proposed metric. Section 4 describes our domain of study based on a characterization of Mars drainage networks. Section 5 reports our experimental analysis, and provides an interpretation of the output clusters. Lastly, Section 6 gives a summary and discusses future work.

2 Preliminaries: Cluster Validation

We assume a dataset of objects, $\mathcal{D} : \{\mathbf{x}_i\}$, where each $\mathbf{x}_i = (a_1, a_2, \dots, a_k)$ 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 will assume each attribute value is a real number, $\mathbf{x}_i \in \mathbb{R}^k$.

A clustering algorithm partitions \mathcal{D} into n mutually exclusive and exhaustive⁴ subsets $\mathcal{K}_1, \mathcal{K}_2, \dots, \mathcal{K}_n$, 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) [2]. Distances are measured according to some predefined metric (e.g., Euclidean distance) over space \mathcal{X} .

We assume the existence of a different mutually exclusive and exhaustive partition of objects, $\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_m$, 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. Our goal is to

⁴ We consider a flat type of clustering (as opposed to hierarchical) where each object is assigned to exactly only cluster.

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 object classifications to validate the set of induced clusters. When a near-optimal match is found we say the clusters have simply recovered a known class structure.

2.1 Metrics Comparing Classes and Clusters

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

Metrics of a statistical nature usually work on 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 on 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. The following statistics have been suggested as metrics of similarity or overlap:

Rand [7]:

$$\frac{\mathcal{E}_{11} + \mathcal{E}_{22}}{\mathcal{E}_{11} + \mathcal{E}_{12} + \mathcal{E}_{21} + \mathcal{E}_{22}} \quad (1)$$

Jaccard [6]:

$$\frac{\mathcal{E}_{11}}{\mathcal{E}_{11} + \mathcal{E}_{12} + \mathcal{E}_{21}} \quad (2)$$

Fowlkes and Mallows [3]:

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

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 [6].

The $m \times n$ Contingency Table

A different approach is to work on a contingency table defined as follows:

Definition 1. A contingency table \mathcal{M} is 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 similarity between \mathcal{C} and \mathcal{K} can be defined in several forms:

Normalized Hamming Distance [4]:

$$\frac{DH_c(\mathcal{M}) + DH_k(\mathcal{M})}{2|\mathcal{D}|} \quad (4)$$

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

$$DH_c(\mathcal{M}) = \sum_i \max_j \mathcal{M}_{ij} \quad (5)$$

$$DH_k(\mathcal{M}) = \sum_j \max_i \mathcal{M}_{ij} \quad (6)$$

Equation 4 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 [1, 11]:

$$H(C|K) = - \sum_i \sum_j \frac{\mathcal{M}_{ij}}{|\mathcal{D}|} \log_2 \frac{\mathcal{M}_{ij}}{\mathcal{M}_j} \quad (7)$$

where \mathcal{M}_j is the marginal sum $\sum_i \mathcal{M}_{ij}$ and lower values are preferred. Equation 7 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 [1].

Limitations

All metrics described above output a numeric value according to the degree of match between \mathcal{C} and \mathcal{K} . In practice, a quantification of the similarity between 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. And even when in principle one could analyze the entries of a contingency matrix to identify clusters having little overlap with existing classes, such information cannot be used in estimating the intersection of the probability models from which the objects were drawn, as it is the case with our parametric approach. We address these issues and our proposed novelty metric next.

3 Assessing the Novelty of New Clusters

We start under the assumption that both clusters and classes can be modelled using a multi-variate Gaussian (i.e., Normal) distribution⁵. In this case the probability density function is completely defined by a mean vector μ and covariance matrix Σ :

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \mu)^t \Sigma^{-1}(\mathbf{x} - \mu)\right] \quad (8)$$

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

Our goal is simply to assess the degree of overlap between a particular class \mathcal{C}_i , modelled as $f_i(\mathbf{x}) : N[\mu_i, \Sigma_i]$, and cluster \mathcal{K}_j , modelled as $f_j(\mathbf{x}) : N[\mu_j, \Sigma_j]$. The lower the degree of overlap the higher the novelty of the cluster. Before explaining our methodology (Section 3.3) we introduce two preliminary metrics.

3.1 The Intersecting Hyper-Volume

A straightforward approach to measure the degree of overlap between $f_i(x)$ and $f_j(x)$, denoted as $\mathcal{O}(f_i(\mathbf{x}), f_j(\mathbf{x}))$, is to calculate the hyper-volume lying at the intersection of both distributions. This can be done by integrating the minimum of both density functions over the whole attribute space:

$$\mathcal{O}(f_i(\mathbf{x}), f_j(\mathbf{x})) = \int_{\mathbf{x}} \min[f_i(\mathbf{x}), f_j(\mathbf{x})] d\mathbf{x} \quad (9)$$

In the extreme case where both distributions have no overlap then $\min[f_i(\mathbf{x}), f_j(\mathbf{x})] = 0$, and hence $\mathcal{O}(f_i(\mathbf{x}), f_j(\mathbf{x})) = 0$. If both distributions are identical or if one distribution is always on top of the other distribution (i.e., if $\forall \mathbf{x} f_i(\mathbf{x}) \geq f_j(\mathbf{x})$ or $\forall \mathbf{x} f_j(\mathbf{x}) \geq f_i(\mathbf{x})$) then $\mathcal{O}(f_i(\mathbf{x}), f_j(\mathbf{x})) = 1$. Although equation 9 can be approximated using numerical methods the computational cost is expensive; the problem soon turns intractable even for moderately low values of n . In practice, a solution to this problem is to assume a form of attribute independence as explained next.

3.2 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 two one-dimensional Gaussian distributions (Figure 1 (left)). We represent the two distributions on attribute A_l , $1 \leq l \leq k$, as $f_i^l(x)$ (corresponding to class \mathcal{C}_i) and $f_j^l(x)$ (corresponding to cluster \mathcal{K}_j). The parameters for these distributions are simply obtained by extracting the corresponding entries on the mean vectors and the diagonal of the covariance matrices.

⁵ This strong assumption is supported by many real domains where data objects can be seen as random disturbances of a prototype data object.

Fig. 1. (left) A measure of the overlap between two distributions; (right) A projection over the difference of the means as a better representation of the separation of the two distributions.

The computation of the overlap of the two distributions, $\mathcal{O}(f_i^l(x), f_j^l(x))$, is now performed over a single dimension and is thus less expensive (equation 9). To combine the degree of overlap over all attributes we adopt a product approximation:

$$\mathcal{O}(f_i(\mathbf{x}), f_j(\mathbf{x})) = \prod_{l=1}^k \mathcal{O}(f_i^l(x), f_j^l(x)) \quad (10)$$

This approach carries some disadvantages. By looking at each attribute independently, two non-overlapping distributions in an n -dimensional space may appear highly overlapped when projected over each attribute. Our challenge lies on finding an efficient approach to estimate $\mathcal{O}(f_i(\mathbf{x}), f_j(\mathbf{x}))$ along a dimension that provides a clear representation of the separation of the two distributions.

3.3 Projecting Over the Difference of the Means

Our proposed solution consists of projecting data objects over a single dimension but in the direction corresponding to the difference of the mean vectors. Specifically, let μ_i be the mean vector of distribution $f_i(\mathbf{x})$ and μ_j be the mean vector of distribution $f_j(\mathbf{x})$; our approach is to project all data objects comprised by class \mathcal{C}_i and cluster \mathcal{K}_j into the new vector

$$\mathbf{w} = \mu_i - \mu_j \quad (11)$$

As illustrated in Figure 1 (right), a projection of data objects over vector \mathbf{w} is often a better indicator of the true overlap between both distributions in n dimensions⁶; it captures the dispersion of data objects precisely along the line cutting through the means. In the extreme case where $\mu_i = \mu_j$, we consider $\mathbf{w} = \mu_i = \mu_j$ which is equivalent to considering the origin as one of the means.

⁶ Alternatively vector \mathbf{w} could be defined as $\mu_j - \mu_i$; both definitions are equally useful.

Once the projection is done, there is no need to work on each attribute separately (as in equation 10); we simply compute the degree of overlap between the projected distributions along vector \mathbf{w} . Thus, our approach efficiently estimates the degree of overlap between two distributions along a single dimension that captures most of the variability⁷ of both class \mathcal{C}_i and cluster \mathcal{K}_j .

Data Projection

To perform the data projection mentioned above we need to compute a scalar dot product

$$x' = \mathbf{w}_0^t \mathbf{x} \quad (12)$$

where \mathbf{x} is an original data point, $\mathbf{w}_0 = \frac{\mathbf{w}}{\|\mathbf{w}\|}$ is a normalized vector such that $\|\mathbf{w}_0\| = 1$, and x' is the (scalar) projection of \mathbf{x} over \mathbf{w}_0 . We project points over the normalized vector \mathbf{w}_0 instead of vector \mathbf{w} simply to give the projection a clear geometrical interpretation (if $\|\mathbf{w}\| \neq 1$ the scale of x' is modified).

We will refer to the projected density functions over \mathbf{w}_0 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}_0 . Let μ be the mean of density function $f(\mathbf{x})$, then the projected parameters are defined as

$$\mu' = \mathbf{w}_0^t \mu \quad \sigma'^2 = \frac{1}{n} \sum (x' - \mu')^2 \quad (13)$$

where μ' and σ'^2 are the projected mean and variance respectively.

In summary, our approach is to quantify the degree of overlap between the two one-dimensional Gaussian distributions $f'_i(x)$ and $f'_j(x)$ obtained after projecting data objects in class \mathcal{C}_i and cluster \mathcal{K}_j along vector \mathbf{w}_0 .

Figure 2 compares our proposed approach with the attribute independence approach (Section 3.2) on an artificial dataset with two Gaussian distributions having the same mean and unit variance. In this experiment there is no cluster novelty, and thus the true overlap (according to equation 9) is one. As shown in Figure 2, both methods tend to stabilize close to one as we increase the sample size, with our proposed approach exhibiting a faster rate of convergence.

A Decomposition of the Degree of Overlap

Until now our measure of overlap has been defined as a function of the intersection of two distributions obtained through a form of data projection. For our purposes we are interested in decomposing the degree of overlap between $f'_i(x)$ and $f'_j(x)$ into two parts:

⁷ This is expected to hold as long as \mathcal{C}_i and \mathcal{K}_j are close to being hyper-circles, or the direction of the principal axes of the hyper-ellipsoids is close to the direction of vector \mathbf{w} .

Fig. 2. A comparison of our approach with the attribute independence approach on an artificial dataset where the true overlap is one. Our approach exhibits a faster rate of convergence.

$$\mathcal{O}(f'_i(x), f'_j(x)) = \int_A f'_j(x) + \int_B f'_i(x) \quad (14)$$

where region A corresponds to all points such that $f'_i \geq f'_j$ and region B corresponds to all points such that $f'_j > f'_i$ (Figure 1 (left)). The decomposition is important to have an understanding of the nature of the overlap. As an example assume cluster \mathcal{K}_j is a proper subset of class \mathcal{C}_i such that $f'_i(x)$ completely covers $f'_j(x)$ (i.e., $\forall x f'_i(x) > f'_j(x)$). In that case all the contribution to the overlap is given by the first integral ($\int_A(\cdot)$). Conversely a cluster covering a class (i.e., $\forall x f'_j(x) > f'_i(x)$) tips all the contribution to the second integral ($\int_B(\cdot)$). We will show later how a correct interpretation of cluster novelty depends on these two numbers (Section 5).

Our implementation of the degree of overlap employs equation 14 instead of equation 9 by computing the extent of regions A and B (i.e., by computing the intersecting points between $f'_i(x)$ and $f'_j(x)$). The output of our algorithm is made of three numbers (equation 14):

1. The total overlap $O_{\mathcal{T}}$
2. The contribution to the overlap by the class ($\int_B(\cdot)$), referred to as $O_{\mathcal{C}}$
3. The contribution to the overlap by the cluster ($\int_A(\cdot)$), referred to as $O_{\mathcal{K}}$

4 Drainage Networks in Mars

We now turn to an area of application where our metric for cluster novelty can be tested. Our study revolves around the morphology of Martian landscapes. The goal in this area is to objectively characterize and categorize Martian landscapes in order to understand their origin. Our study uses a dataset characterizing different regions on Mars from the perspective of how they drain. Martian topography based on MOLA data is used to represent landscapes as a series of drainage basins, regardless of the historical presence

or absence of actual fluid flow [10]. A drainage network, the part of a basin where the flow is concentrated, is computationally delineated from the basin. Such network has a fractal structure which is described in terms of probability distribution functions of various drainage quantities. Following the method described in [10], the morphology of each network can be encapsulated in a network descriptor or vector of four numbers $\mathbf{x} = (\tau, \gamma, \beta, \rho)$. Briefly, τ , γ , and β are attributes that characterize distributions of contributing areas, lengths of main streams, and dissipated energy, respectively; parameter ρ measures the spatial uniformity of drainage. The network descriptor offers an abstract but very compact characterization of a drainage network.

Our dataset consist of 386 data objects derived from Martian landscapes with a wide range of latitudes and elevations. Our study aims at determining if our characterization of Martian landscapes clusters into natural groups and if there can be a clear interpretation of such clusters. We compare our clusters to a known traditional and descriptive characterization of the Martian surface that divides it into a number of classes called geological units [8]. Division into geological units is based on terrain texture, its geological structure, its age, and its stratigraphy. All this attributes are determined from visual inspection of imagery data. For example unit Hr is described as having "moderately cratered surface, marked by long, linear or sinuous ridges"; unit Np1 is "highland terrain with high density of craters." Our objects are extracted from surfaces belonging to 16 different geological units representing three major Martian epochs: Noachian, Hesperian, and Amazonian. On the other hand, our clusters group together terrains based on similar drainage patterns. These patterns are obtained from digital topography using a computer algorithm. There is no a priori no clear relation between the two classifications.

5 Empirical Study

We divide our empirical study into two steps: 1) an assessment of the similarity (dissimilarity) of a set of clusters of Mars landscapes with the set of known geological units and 2) an interpretation of the clusters based on hydrological aspects of landscape morphology. We look at each step in turn.

5.1 Cluster Generation

The probabilistic clustering algorithm corresponding to our experiments follows the Expectation Maximization (EM) technique [5]. It groups records into clusters by modelling each cluster through a probability density function. Each record in the dataset has a probability of class membership and is assigned to the cluster with highest posterior probability. The number of clusters is estimated using cross-validation; the algorithm is part of the WEKA machine-learning tool [12].

Applying the EM clustering algorithm directly over the drainage network dataset results in a partition corresponding to nine different clusters. The next step is to assess the degree of match between our nine clusters and the sixteen Martian geological units following our proposed approach (Section 3.3).

Table 1. A measure of the degree of overlap between clusters and classes in the context of Martian topography.

Clusters	Geological Units				
	Most Similar	2nd	3rd	4rd	5th
C1	0.786 Hr (0.309, 0.476)	0.768 Nplr (0.216, 0.552)	0.713 Npld (0.192, 0.522)	0.686 Hnu (0.455, 0.232)	0.669 Npl1 (0.171, 0.498)
C2	0.846 Aoa (0.594, 0.252)	0.533 Aps (0.218, 0.316)	0.438 Hnu (0.141, 0.297)	0.406 Hr (0.082, 0.324)	0.384 Npl1 (0.076, 0.308)
C3	0.372 Hvk (0.127, 0.245)	0.362 Hnu (0.070, 0.293)	0.341 Apk (0.064, 0.277)	0.2485 Npld (0.041, 0.208)	0.158 Nh1 (0.023, 0.136)
C4	0.818 Aps (0.260, 0.559)	0.647 Nplr (0.162, 0.485)	0.591 Nh1 (0.142, 0.449)	0.577 Hpl3 (0.139, 0.439)	0.566 Hnu (0.133, 0.433)
C5	0.723 Hh3 (0.248, 0.476)	0.431 Nh1 (0.093, 0.338)	0.353 Hnu (0.069, 0.285)	0.345 Nplr (0.168, 0.177)	0.343 Aps (0.102, 0.241)
C6	0.638 Apk (0.263, 0.375)	0.450 Npld (0.099, 0.351)	0.346 Nh1 (0.066, 0.280)	0.329 Hnu (0.061, 0.268)	0.302 Npl1 (0.054, 0.248)
C7	0.784 Hr (0.246, 0.538)	0.738 Nh1 (0.364, 0.374)	0.696 Nplr (0.182, 0.514)	0.663 Hnu (0.377, 0.287)	0.513 Apk (0.209, 0.304)
C8	0.938 Hpl3 (0.357, 0.581)	0.851 Npl1 (0.290, 0.560)	0.849 Hnu (0.603, 0.247)	0.837 Npld (0.596, 0.241)	0.821 Nplr (0.578, 0.244)
C9	0.488 Nh1 (0.255, 0.233)	0.438 Hh3 (0.113, 0.325)	0.389 Hr (0.232, 0.158)	0.369 Npl1 (0.072, 0.298)	0.278 Nplr (0.048, 0.231)

5.2 Comparing Clusters to Geological Units

Table 1 shows our results. The first column corresponds to the nine clusters obtained over the drainage network dataset. For each row, the second column corresponds to the class (i.e., geological unit) with highest overlap to that cluster, the third column corresponds to the class with the second highest overlap, and so on. We report on the five classes with highest overlap for each cluster. On each entry we report the total degree of overlap between the cluster and the class (O_T) and the corresponding geological unit; within parentheses we show the contribution of the cluster and class to the overlap (O_C, O_K).

A first glance at Table 1 may indicate a relatively high overlap between clusters and at least some classes. In some cases this can be explained by looking at the two components of the overlap separately. For example, in some clusters, a high degree of overlap is an artifact of their small size (e.g., cluster C8 with unit Hpl3—a hint is the relatively large value of the second component in the overlap, O_K , showing how the unit covers a large portion of the cluster). In other cases, clusters such as C6, C7, and C9, contain many objects and still display a sizable overlap with selected geological units. Using expert knowledge about hydrological properties of clusters and geological properties of classes we may assess whether that overlap is evidence of any significant correlation or if the clusters point to a novel classification of Mars terrains.

Fig. 3. Four martian terrains from two different geological units and belonging to two different clusters. Drainage networks are drawn on top of the terrain. It is easy to see similarity based on geological unit, but also similarity based on our clustering.

5.3 Interpretation of Results

The apparent similarity between some clusters and geological units can be explained by observing that basins of all shapes can form in any geological setting. Figure 3 illustrates this point. Four terrains are shown in a 2×2 matrix arrangement. Terrains in the same row belong to the same geological unit, terrains in the same column belong to the same cluster. It is easy to see similarity based on geological unit (look at the distribution of craters across rows), but using drainage networks drawn on top of the terrain, it is also easy to see similarity based on our clustering (look at the shape of the drainage networks across columns).

Cluster C7 describes terrains characterized by drainage basins with widths about the same as their longitudinal extents, these are terrains made of "squared basins". In contrast, cluster C9 groups terrains with narrow basins. The higher overlap of C7 with Hr (0.784) compared to cluster C9 with Hr (0.389) suggests that texture of Hr terrain is such that "squared basins" are preferentially formed (look at the difference in the second component of the overlap). Similar conclusion is also true about the Apk terrain. The most populous cluster, C6, groups terrains that are neither too narrow, nor too squared; it overlaps moderately with a number of geological units. Our interpretation is that *there is no correlation between the classifications based on geology and hydrology*. The texture of all the Martian terrain is such that "squared basins" are simply a common form. This is why a cluster of such objects is the most populous, and this is why it overlaps with

many geological units. Other basin shapes are rarer and thus show smaller overlap with geological units. Some clusters have only a few objects which may happen to belong to particular classes; that results in a false indication of high overlap.

6 Summary and Conclusions

Assessing the degree of match between the partitions induced by a clustering algorithm against those induced by an external classification is normally done through a single numeric response. In this paper we introduce a different approach that quantifies the degree of overlap between two one-dimensional Gaussian distributions obtained after projecting data objects along the vector that cuts through the means.

We test our approach on Martian landscapes by comparing each induced cluster with a set of classes known as geological units. The apparent high overlap between some classes and clusters can be explained by attending to the different nature of the partitions induced by both classifications. Whereas our clusters provide a hydrological view of terrains, the set of existing classes provide a geological view. Both views tend to share terrains whereas in fact there is no correlation among them. In addition, our decomposition of the degree of overlap (Section 3.3) has proved instrumental in giving an interpretation to the similarity (dissimilarity) between clusters and classes.

Acknowledgments

Thanks to the Lunar and Planetary Institute for facilitating data on Martian landscapes.

References

1. Dom Byron: An Information-Theoretic External Cluster-Validity Measure. Research Report, IBM T.J. Watson Research Center RJ 10219 (2001)
2. Duda R. O., Hart P. E., Stork D. G.: Pattern Classification. John Wiley Ed. 2nd Edition (2001)
3. Fowlkes E., Mallows C.: A Method for Comparing Two Hierarchical Clusterings. *Journal of American Statistical Association*, 78 pp. 553–569 (1983).
4. Kanungo T., Dom B., Niblack W., Steele D.: A Fast Algorithm for MDL-Based Multi-Band Image Segmentation. *Image Technology*, Jorge Sanz (ed.) Springer-Verlag (1996).
5. McLachlan G., Krishnan T.: *The EM Algorithm and Extensions*. John Wiley and Sons (1997).
6. Milligan G. W., Soon S. C., Sokol L. M.: The Effect of Cluster Size, Dimensionality, and the Number of Clusters on Recovery of True Cluster Structure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 5, No. 1 pp. 40–47 (1983).
7. Rand W. M.: Objective Criterion for Evaluation of Clustering Methods. *Journal of American Statistical Association*, 66 pp. 846–851 (1971).
8. Scott D.H., Carr M.H.: Geological Map of Mars. U.S.G.S. Misc Geol. Inv. Map I-1093 (1977).
9. Smith, D.E., et al.: Mars Orbiter Laser Altimeter: Experiment summary after the first year of global mapping of Mars. *J. Geophys. Res.*, Vol. 106, 23,689–23,722 (2001).
10. Stepinski T., Marinova M. M., McGovern P.J., Clifford S. M.: Fractal Analysis of Drainage Basins on Mars. *Geophysical Research Letters*, Vol. 29, No. 8 (2002).
11. Vaithyanathan S., Dom B.: Model Selection in Unsupervised Learning with Applications to Document Clustering. *Proceedings of the Sixteenth International Conference on Machine Learning*, Stanford University, CA (2000).
12. Witten I. H., Frank E.: *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Academic Press, London U.K (2000).