# Cluster Validation

**Ricardo Vilalta**
*University of Houston, USA*

**Tomasz Stepinski**
*Lunar and Planetary Institute, USA*

## INTRODUCTION

Spacecrafts orbiting a selected suite of planets and moons of our solar system are continuously sending long sequences of data back to Earth. The availability of such data provides an opportunity to invoke tools from machine learning and pattern recognition to extract patterns that can help to understand geological processes shaping planetary surfaces. Due to the marked interest of the scientific community on this particular planet, we base our current discussion on Mars, where there are presently three spacecrafts in orbit (e.g., NASA's Mars Odyssey Orbiter, Mars Reconnaissance Orbiter, ESA's Mars Express). Despite the abundance of available data describing Martian surface, only a small fraction of the data is being analyzed in detail because current techniques for data analysis of planetary surfaces rely on a simple visual inspection and descriptive characterization of surface landforms (Wilhelms, 1990).

The demand for automated analysis of Mars surface has prompted the use of machine learning and pattern recognition tools to generate geomorphic maps, which are thematic maps of landforms (or topographical expressions). Examples of landforms are craters, valley networks, hills, basins, etc. Machine learning can play a vital role in automating the process of geomorphic mapping. A learning system can be employed to either fully automate the process of discovering meaningful landform classes using *clustering* techniques; or it can be used instead to predict the class of unlabeled landforms (after an expert has manually labeled a representative sample of the landforms) using *classification* techniques. The impact of these techniques on the analysis of Mars topography can be of immense value due to the sheer size of the Martian surface that remains unmapped.

While it is now clear that machine learning can greatly help in automating the detailed analysis of Mars' surface (Stepinski et al., 2007; Stepinski et al., 2006; Bue and Stepinski, 2006; Stepinski and Vilalta, 2005), an interesting problem, however, arises when an automated data analysis has produced a novel classification of a specific site's landforms. The problem lies on the interpretation of this new classification as compared to traditionally derived classifications generated through visual inspection by domain experts. Is the new classification novel in all senses? Is the new classification only partially novel, with many landforms matching existing classifications? This article discusses how to assess the value of clusters generated by machine learning tools as applied to the analysis of Mars' surface.

## BACKGROUND ON CLUSTER VALIDATION

We narrow our discussion to patterns in the form of clusters as produced by a clustering algorithm (a form of unsupervised learning). 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). The goal is to discover how data objects gather into natural groups (Duda at el., 2001; Bishop, 2006). The application of clustering algorithms can be followed by a post-processing step, also known as cluster validation; this step is commonly employed to assess the quality and meaning of the resulting clusters (Theodoridis and Koutroumbas, 2003).

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 (Krishnapuran et al., 1995; Theodoridis and Koutroumbas, 2003). As an example, one can measure the quality of the resulting clusters by assessing the degree of compactness of the clusters, or the degree of separation between clusters.

On the other hand, 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*. In the context of planetary science, for example, a collection of sites on a planet constitutes a set of objects that are classified manually by domain experts (geologists) on the basis of their geological properties. In the case of planet Mars, the resultant division of sites into the so-called *geological units* represents an external classification. A clustering algorithm that is invoked to group sites into different clusters can be compared to the existing set of geological units to determine the novelty of the resulting clusters.

Current approaches to external cluster validation are 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 (Dom, 2001). Such narrow assumption precludes alternative interpretations; in some scenarios high-quality clusters are considered novel if they do not resemble existing classes. After all, a large separation between clusters and classes can serve as clear evidence of cluster novelty (Cheeseman and Stutz, 1996); on the other hand, finding clusters resembling existing classes serves to confirm existing theories of data distributions. 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 summary, most traditional metrics for external cluster validation output a single value indicating the degree of match between the partition induced by the known classes and the one induced by the clusters. We claim this is the wrong approach to validate patterns output by a data-analysis technique. By averaging the degree of match across all classes and clusters, traditional metrics fail to identify the potential value of individual clusters.

## CLUSTER VALIDATION IN MACHINE LEARNING

The question of how to validate clusters appropriately without running the risk of missing crucial information can be answered by avoiding any form of averaging or smoothing approach; one should refrain from computing an average of the degree of cluster similarity with respect to external classes. Instead, we claim, one should compute the distance between each individual cluster and its most similar external class; such comparison can then be used by the domain expert for an informed cluster-quality assessment.

### Traditional Approaches to Cluster Validation

More formally, the problem of assessing the degree of match between the set **C** of predefined classes and the set **K** of new clusters is traditionally performed by computing a metric where high values indicate a high similarity between classes and clusters. For example, one type of statistical metric is defined in terms of a **2x2** table where each entry $\mathbf{E}_{ij}$, i,j Є {1,2}, counts the number of object pairs that agree or disagree with the class and cluster to which they belong; $\mathbf{E}_{11}$ corresponds to the number of object pairs that belong to the same class and cluster, $\mathbf{E}_{12}$ corresponds to same class and different cluster, $\mathbf{E}_{21}$ corresponds to different class and same cluster, and $\mathbf{E}_{22}$ corresponds to different class and different cluster. Entries along the diagonal denote the number of object pairs contributing to high similarity between classes and clusters, whereas elements outside the diagonal contribute to a high degree of dissimilarity. A common family of statistics used as metrics simply average correctly classified class-cluster pairs by a function of all possible pairs. A popular similarity metric is Rand's metric (Theodoridis and Koutroumbas, 2003):

$(E_{11} + E_{22}) / (E_{11} + E_{12} + E_{21} + E_{22})$

Other metrics are defined as follows:

Jaccard:

$\mathbf{E_{11} / (E_{11} + E_{12} + E_{21})}$

Fowlkes and Mallows:

$\mathbf{E_{11} / [(E_{11} + E_{12}) (E_{21} + E_{22})]^{1/2}}$

A different approach is to work on a contingency table **M**, defined as a matrix of size **m**x**n** where each row corresponds to an external class and each column to a cluster. An entry $\mathbf{M}_{\mathbf{ij}}$ indicates the number of objects covered by class $\mathbf{C}_{\mathbf{i}}$ and cluster $\mathbf{K}_{\mathbf{j}}$. Using **M**, the similarity between **C** and **K** can be quantified into a single number in several forms (Kanungo et al., 1996; Vaithyanathan and Dom, 2000).

## Limitations of Current Metrics

In practice, a quantification of the similarity between sets of classes and clusters is of limited value; we claim any potential discovery provided by the clustering algorithm is only identifiable by analyzing the meaning of each cluster individually. As an illustration, assume two clusters that bear a strong resemblance with two real classes, with a small novel third cluster bearing no resemblance to any class. Averaging the similarity between clusters and classes altogether disregards the potential discovery carried by the third cluster. If the third cluster is small enough, most metrics would indicate a high degree of class-cluster similarity.

In addition, 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 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; probabilistic expectations can differ significantly from actual counts because the probabilistic model introduces substantial a priori information.

## Proposed Approach to Cluster Validation

We show our approach to cluster validation in the context of the analysis of Mars' surface (Vilalta et al., 2007). The current qualitative means of classifying Martian surfaces is by assigning each site to what is called a *geological unit*. One shortcoming of such classification is that it cannot be automated because it is normally assigned subjectively by a domain expert. On the other hand, an automated classification of Martian surfaces is possible using digital topography data. Martian topography data is currently available from the Mars Orbiter Laser Altimeter (MOLA) instrument (Smith et al., 2003). This data can be used to construct a digital elevation model (DEM) of a site on Mars' surface. The DEM is a regular grid of cells with assigned elevation values. Such grid can be processed to generate a dataset of feature vectors (each vector component stands as a topographic feature). The resulting training data set can be used as input to a clustering algorithm.

If the clustering algorithm produces **N** clusters, and there exists **M** externals classes (in our case study **M=16** sixteen Martian geological units), we advocate validating the quality of the clusters by computing the degree of overlap between each cluster and each of the existing external classes. In essence, one can calculate an **NxM** matrix of distances between the clusters and classes. This approach to pattern validation enables us to assess the value of each cluster independently of the rest, and can lead to important discoveries.

A practical implementation for this type of validation is as follows. One can model each cluster and each class as a multivariate Gaussian distribution; the degree of separation between both distributions can then be computed using an information-theoretic measure known as relative entropy or Kullback-Leibler distance (Cover and Thomas, 2006). The separation of two distributions can be simplified if it is done along a single dimension that captures most of the data variability (Vilalta et al., 2007). One possibility is to project all data objects over the vector that lies orthogonal to the hyper-plane that maximizes the separation between cluster and class, for example, by using Fisher's Linear Discriminant (Duda, et al., 2001). The resulting degree of overlap can be used as a measure of the similarity of class and cluster. As mentioned before, this approach enables us to assess clusters individually. In some cases a large separation (low overlap) may indicate domain novelty, whereas high overlap or similarity may serve to reinforce current classification schemes.

## A CASE STUDY IN PLANETARY SCIENCE

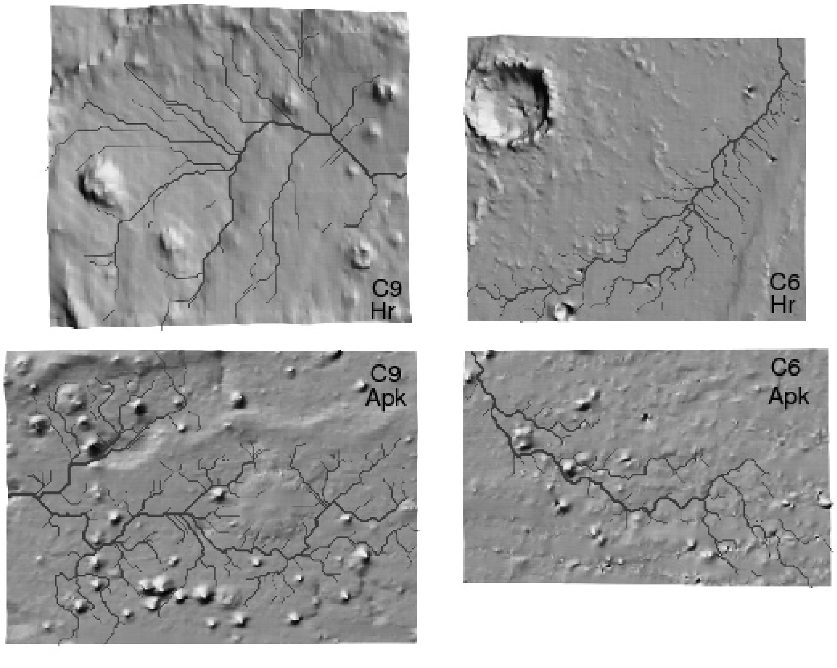
In the analysis of Mars' surface, our clustering algorithm produced **N=9** clusters. Using the method for external cluster assessment explained above, we were able to determine that partitioning the dataset of Martian sites on the basis of the resulting clusters produced a novel classification that does not match the traditional classification based on (**M=16**) geological units. We could conclude this by analyzing each cluster individually, observing no close resemblance with existing classes (in addition our methodology indicates which clusters are more dissimilar than others when compared to the set of classes).

Figure 1 shows an example of the difference between patterns found by clusters and those pertaining to more traditional geomorphic attributes. Four Martian surfaces are shown in a **2x2** matrix arrangement. Surfaces in the same row belong to the same geological unit, whereas surfaces in the same column belong to the same cluster. The top two sites show two surfaces from a 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 exhibit very different shape for their drainage (blue river-like features). The bottom two sites show two surfaces from a geological unit described as "smooth plain with conical hills or knobs." Again, it is easy to see the similarity between these two surfaces based on that description. Nevertheless, the two terrains have drainage with markedly different character. On the basis of the cluster distinction, these four surfaces could be divided vertically instead of horizontally. Such division corresponds to our cluster partition where the drainage is similar for sites within the same cluster. Our methodology facilitates this type of comparisons because clusters are compared individually to similar classes. Domain experts can then provide a scientific explanation of the new data categorization by focusing on particular differences or similarities between specific class-cluster pairs.

## FUTURE TRENDS

Future trends include assessing the value of clusters obtained with alternative clustering algorithms (other than probabilistic algorithms). Another trend is to devise modeling techniques for the external class distribution (e.g., as a mixture of Gaussian 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.

*Figure 1. Four Martian surfaces that belong to two different geological units (rows), and two different clusters (columns). Drainage networks are drawn on top of the surfaces.*



## CONCLUSION

Cluster validation looks for methods to assess the value of patterns output by machine learning techniques. In the context of unsupervised learning or clustering, data objects are grouped into new categories 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 document we criticize common approaches to pattern validation where the mechanism consists of computing an average of the degree of similarity between clusters and classes. Instead we advocate an approach to external cluster assessment based on modeling each cluster and class as a probabilistic distribution; the degree of separation between both distributions can then be measured using an information-theoretic approach (e.g., relative entropy or Kullback-Leibler distance). By looking at each cluster individually, one can assess the degree of novelty (large separation to other classes) of each cluster, or instead the degree of validation (close resemblance to other classes) provided by the same cluster.

## REFERENCES

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. New York, NY: Springer.

Bue, B. D., & Stepinski, T. F. (2006). Automated Classification of Landforms on Mars, *Computers & Geoscience*, 32(5), 604-614.

Cheeseman, P., & Stutz, J. (1996). Bayesian Classification (AutoClass): Theory and Results. In Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (Ed.), *Advances in Knowledge Discovery and Data Mining,* AAAI Press/MIT Press.

Cover, T. M., & Thomas, J. A. (2006). *Elements of Information Theory*. John Wiley & Sons.

Dom, B. (2001). *An Information-Theoretic External Cluster-Validity Measure* (Tech. Research Rep. No. 10219). IBM T.J. Watson Research Center.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern Classification*. Canada: John Wiley, 2nd Edition.

Kanungo, T., Dom., B., Niblack, W., & Steele, D. (1996). A Fast Algorithm for MDL-based Multi-Band Image Segmentation. In: Sanz, J. (ed) Image Technology. Springer-Verlag, Berlin.

Krishnapuran, R., Frigui, H., & Nasraoui, O. (1995). Fussy and Possibilistic Shell Clustering Algorithms and Their Application to Boundary Detection and Surface Approximation, Part II, *IEEE Transactions on Fuzzy Systems*, 3(1), 44-60.

Smith, D., Neumann, G., Arvidson, R. E., Guinness, E. A., & Slavney, S. (2003). Mars Global Surveyor Laser Altimeter Mission Experiment Gridded Data Record, *NASA Planetary Data System, MGS-M-MOLA-5-MEGDR-L3-V1.0.*

Stepinski, T. F., Ghosh, S., & Vilalta, R. (2007). Machine Learning for Automatic Mapping of Planetary Surfaces. *Nineteenth Innovative Applications of Artificial Intelligence Conference*.

Stepinski, T. F., Ghosh, S., & Vilalta, R. (2006). Automatic Recognition of Landforms on Mars Using Terrain Segmentation and Classification. *International Conference on Discovery Science* (pp. 255-266). LNAI 4265.

Stepinski, T. F., & Vilalta, R. (2005). Digital Topography Models for Martian Surfaces, *IEEE Geoscience and Remote Sensing Letters*, 2(3), 260-264.

Theodoridis, S. & Koutroumbas, K. (2003). *Pattern Recognition.* San Diego, CA: Academic Press.

Vaithyanathan, S., & Dom., B. (2000). Model Selection in Unsupervised Learning with Applications to Document Clustering. Proceedings of the 6th International Conference on Machine Learning, Stanford University, CA.

Vilalta, R., Stepinski, T., & Achari, M. (2007). An Efficient Approach to External Cluster Assessment with an Application to Martian Topography. *Data Mining and Knowledge Discovery Journal*, 14, 1-23.

Wilhelms, D. E. (1990). Geologic Mapping. In: Greeley, R., Batson, R. (Eds.), *Planetary Mapping.* Cambridge University Press, Cambridge, UK, pp 209-244.

## KEY TERMS

**Cluster Validation:** A post-processing step after clustering used to assess the value of the resulting clusters.

**Digital Elevation Model:** A digital elevation model (DEM) is a regular grid of cells with assigned elevation values. It characterizes a particular site based on the shape of the terrain.

**External Cluster Validation:** Cluster validation is called *external* if the validation is performed by gathering statistics comparing the induced clusters against an external and independent classification of objects.

**Fisher's Linear Discriminant:** Fisher's linear discriminant finds a hyperplane that separates two data clusters by searching for a normal vector **w** that maximizes the separation between clusters when the data is projected over **w**.

**Internal Cluster Validation:** Cluster validation is called *internal* when statistics are devised to capture the quality of the induced clusters using the available data objects only.

**MOLA Instrument:** MOLA is the Mars Orbiter Laser Altimeter; it is an instrument attached to the Mars Global Surveyor spacecraft sent to Mars in 1996.

**Relative Entropy:** The relative entropy or Kullback-Leibler distance between two probability mass functions **p(x)** and **q(x)** is defined as $\mathbf{D(p \mid\mid q) = \Sigma_x p(x) \log [p(x)/q(x)]}$ (Cover & Thomas, 2006)**.**