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 evaluating 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 $\mathbf{E}_{ij}$, $i,j \in \{1,2\}$, 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):

$$(\mathbf{E}_{11} + \mathbf{E}_{22}) / (\mathbf{E}_{11} + \mathbf{E}_{12} + \mathbf{E}_{21} + \mathbf{E}_{22})$$

Other metrics are defined as follows:

Jaccard:

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

Fowlkes and Mallows:

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