The Importance of Informative Feature Representations

Ricardo Vilalta

1 Feature Representation, Bias, Variance, and Irreducible Error

Typical issues under consideration when selecting or designing a classification algorithm are the bias and variance components of error induced by the algorithm [1]. For example, one may choose a simple algorithm (e.g., linear combination of feature values, Naive Bayes, single logical rules, etc.) and draw a hypothesis from a small family of functions; the poor repertoire of functions may produce high bias (the best function may be far from the target function) but low variance (because of the sensitivity on local data irregularities). The alternative is to increase the degree of complexity by drawing a hypothesis from a large class of functions (e.g., neural networks with a large number of hidden units); here the hypothesis exhibits flexible decision boundaries (low bias) but becomes sensitive to small variations in the data (high variance).

A less explored –but perhaps more critical issue– is that of the feature representation, which can be the cause of a third component of error known as Bayes (irreducible) error. This occurs when the feature representation leads to class overlap. While bias and variance can be traded off by varying the classification strategy, Bayes error remains immutable as soon as the feature representation is fixed. The importance of high quality features is crucial to attain accurate predictions and cannot be over-emphasized [2]. High quality features convey much information about the problem; in this case, even a simple hypothesis suffices to produce good results. In contrast, low quality features complicate the classification process. Features can bear poor correlation with the class, or interact in many ways, which calls for additional steps to discover important feature combinations.

Dept. of Computer Science, University of Houston, 501 Philip G. Hoffman, Houston TX 77204-3010, e-mail: vilalta@cs.uh.edu

2 A Commentary on Sunspot Classification

The paper by Stenning, et. al. (this volume) entitled "Morphological Image Analysis and its Application to Sunspot Classification" describes an interesting approach to sunspot classification using techniques from mathematical morphology and image processing. The authors describe a well-thought set of techniques to differentiate among four classes, following the Mount-Wilson classification scheme. Such classes vary according to the shape and distribution of magnetic flux in sunspots groups. It is clear from the paper that the task of extracting relevant features to differentiate among such classes is extremely difficult. The distribution of positive and negative magnetic polarities extracted from the magnetogram can exhibit multiple configurations, which makes it very difficult to point to the right class precisely. The paper gives a hint at the strong challenge of acquiring additional features to improve on accuracy performance (currently reported at around 58% on a testing set using random forests).

The problem of finding relevant features in images with spatial content appears in many other scientific domains. We have found that one key element to discover relevant features in these problems is to look for contextual information according to the precise nature of the classes under analysis. One particular domain is that of automatic classification of landforms on Mars, described next.

An Analogous Problem in Landform Classification on Mars

We follow our discussion with a brief description of a pattern recognition tool for mapping landforms on Mars [3, 4] that receives as input a DEM (Digital Elevation Map). It uses the values of elevations stored in the DEM to calculate additional geomorphometric features; we use the following cell-based features: slope, curvature, and flooding adjustment. At the end of our feature generation process we have a 3-dimensional feature vector assigned to each cell in the raster. The raster is then segmented into spatially single-connected, feature vectors. After segmentation, the raster consists of a number of spatial patches; these patches are the objects of final classification based on six possible classes: inter-crater plateau, crater floor, convex crater wall, concave crater wall, convex ridge, and concave ridge.

Figure 1 shows two images of Mars. Figure 1-top describes the "ground-truth", where all segments have been correctly classified by an expert, with class labels displayed on the right side. Figure 1-bottom shows the result of automatically classifying that region of Mars using a small amount of segments for training, and using the rest for testing. The problem with this task is that there are semantically different landforms that display similar or even identical landscape elements. This is difficult because it requires domain knowledge of Martian topology as regards to structural shape. An example of two distinct landforms consisting of very similar landscape elements is the case of concave crater walls and concave ridges. Both landscape elements are rim-like surfaces, the difference is that in the case of the

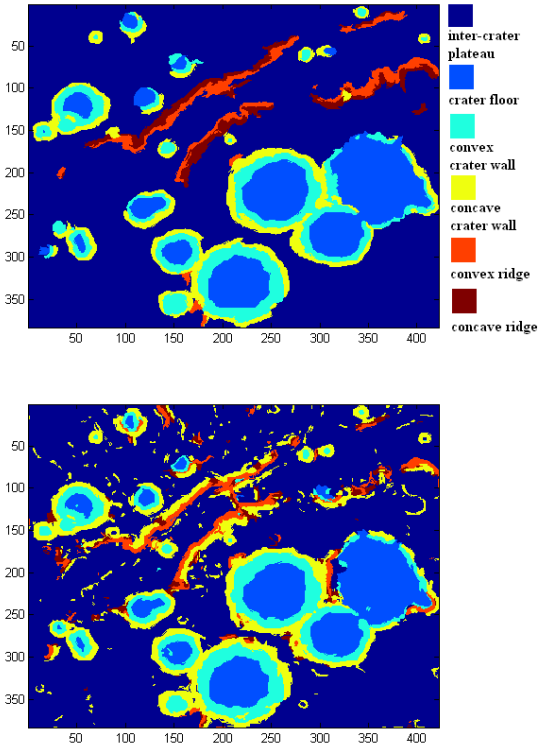

Fig. 1 Top: A color-labeling of a site on Mars (Tisia Site) with perfect landform classification. Bottom: The approximation made by a learning algorithm (Support Vector Machines) using local features.

crater, a collection of rim structures forms a circle-like landform, and, in the case of the ridge, it forms a linear-like structure. This is again a problem where the feature representation is crucial to attain good results. In the case of Mars, additional features capturing the shape and global distribution of segments on each landform are necessary to overcome the irreducible error that comes from class overlapping.

References

1. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York (2009)
2. Liu, H., Motoda, H.: Computational Methods of Feature Selection. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series (2007).
3. Stepinski T., Vilalta R.: Machine Learning Tools for Geomorphic Mapping of Planetary Surfaces. In: Zhang Y. (eds.) Machine Learning, pp. 251-266. In-Tech (2010)

4. Ghosh S., Stepinski T., Vilalta R.: Automatic Annotation of Planetary Surfaces with Geomorphic Labels. IEEE Transactions on Geoscience and Remote Sensing. **48**, 175–185 (2009)