14

Machine Learning Tools for Geomorphic Mapping of Planetary Surfaces

Tomasz F. Stepinski[1] and Ricardo Vilalta[2]
[1]*Lunar and Planetary Institute, Houston, TX*
[2]*University of Houston, Houston, TX*
USA

1. Introduction

Terrain topography carries information that is fundamental for geomorphic modeling and, ultimately, for understanding geologic processes responsible for land-surface form. Classification of terrain is tantamount to organizing the expressions (features) of terrain topography into landforms (classes) – patches of topography having similar characteristics and commonly recognizable semantic labels. Because of the spatial character of topographic data, such classification is referred to as geomorphic mapping. The result of classification of an entire landscape scene into a set of mutually exclusive and exhaustive landform classes is referred to as a geomorphic map.

Geomorphic mapping of terrestrial and planetary surfaces has been done traditionally via visual interpretation of images (Wilhelms, 1990; Tanaka, 1994). This manual method is slow, labor intensive, and suffers from subjectivity. Presently, remote sensing instruments onboard spacecrafts are providing increasingly large volumes of topographic data related to Earth as well as surfaces of other planets. This data rich environment challenges the ability of the geosciences community to turn a significant portion of all collected data into products (like, for example, geomorphic maps) that could be utilized in research. Simply put, advances in geomorphic mapping have not kept up with advances in data collection. If left to manual mapping, the percentage of planetary surfaces mapped to the level of detail permitted by an increased resolution of newly collected data will continue to drop precipitously. In order to prevent this decline in rate of data to map conversion, it is necessary to automate the mapping process. Fortunately, the surface properties that distinguish between different landforms can be described quantitatively by a set of numerical measures called terrain attributes which are derived from a digital elevation model (DEM) of the surface. This opens the opportunity for automation of the mapping process. The topic of auto-mapping landforms from topography has received some attention in the geosciences literature; however, such approaches rely frequently on hand-made rules for classification and are designed exclusively for terrestrial applications.

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. We refer to the two techniques as unsupervised and supervised learning, respectively. Unsupervised learning techniques are applicable in cases of *exploratory* mapping, where no prior knowledge about the surface exists and both landform types and their spatial presence need to be derived by an algorithm. Exploratory mapping finds application in expediting creation of geologic maps in planetary science context where surfaces are still being explored and landform classes are not yet defined. Supervised learning techniques are applicable in cases of *exploitation* mapping when only the spatial presence of a priori defined landform classes is required. Exploitation mapping finds application in creating final geomorphic maps for terrestrial and planetary sites for which constituting landform classes are known a priori.
In developing machine learning-based mapping tools we face a number of design choices, starting from the selection of a basic unit of surface, through the choice of features (terrain attributes), to a pick of an appropriate machine learning technique. The fundamental problem is to design a technique that results in a map that has information content and visual esthetics similar to those found in manually drawn maps; such outcome ensures a large impact in the geosciences community and, consequently, has the greatest scientific value.

2. Preliminaries and previous work

All our tools are based on topographic data which provides a fundamental description of a surface and is well-suited for automated mapping. Topographic data is available as a grid-based DEM, a raster that stores site's elevation value at each pixel in a corresponding raster node. All features used by our machine learning-based mapping tools are derived from the DEM. These features are divided into at-pixel features and area statistics features (Evans, 1998). At-pixel features, except for elevation itself, require a small region or neighborhood around the pixel to calculate their values. Area statistics features depend on the range or distribution of values within the selected, larger neighborhood.
Previously published methods for auto-mapping of landforms can be divided into those that utilize machine learning and those that don't (Gallant et al, 2005; Dragut and Blaschke, 2006; van Asselen and Seijmonsbergen, 2006; Iwahashi and Pike, 2007). Machine learning-based methods can be further grouped into those using unsupervised learning techniques (Irvin et al, 1997; Burrough et al, 2000; Adediran et al, 2004; Stepinski and Vilalta, 2005; Bue and Stepinski, 2006; Ehsani and Quiel, 2008;) and those using supervised learning (Brown et al, 1998; Hengl and Rossiter, 2003; Prima et al, 2006; Stepinski et al, 2006; Stepinski et al, 2007). Moreover, all methods can be grouped into pixel-based methods (Irvin et al, 1997; Brown et al, 1998; Burrough et al, 2000; Hengl and Rossiter, 2003; Adediran et al, 2004; Stepinski and Vilalta, 2005; Bue and Stepinski, 2006; Prima et al, 2006; Iwahashi and Pike, 2007; Ehsani and Quiel, 2008), where an algorithm assigns landform label for each pixel in a DEM separately, and segmentation-based methods (Dragut and Blaschke, 2006; van Asselen and Seijmonsbergen, 2006; Stepinski et al, 2006; Stepinski et al, 2007; Ghosh et al, 2009), where an algorithm assigns landform labels to multi-pixel but attribute-homogeneous segments of the landscape. Fig. 1. illustrates the conceptual difference between pixel-based and segmentation-based approaches. Proposed methods differ broadly in classification algorithms and feature selection. We claim that a segmentation-based approach utilizing

supervised or unsupervised machine learning has the potential to generate maps most comparable to manual maps, and thus most useful. Consequently, our own recent efforts have concentrated on such approaches. Here we discuss the spectrum of tools that we have developed for both exploratory and exploitation purposes. We discuss the tools for exploratory mapping that are both, pixel-based and segmentation-based. We also discuss the tools for exploitation mapping, which are exclusively segmentation-based. Our work focuses on mapping the planet Mars, because Mars is the only planet besides Earth for which global topographic data is currently available (Smith et al, 2003). However, the tools are applicable to mapping terrestrial landmass for which a global, high resolution DEM is available. Moreover, these tools are also applicable to mapping the surfaces of planet Mercury and the Moon once the DEMs for these planets become available (Krishna et al, 2009; Araki et al, 2009) in the near future. The resolution of DEMs of planetary surfaces is coarser than the resolution of terrestrial DEMs. This presents unique challenges for auto-mapping their surfaces as the coarse resolution excludes the direct use of area statistics features decreasing the number of features available to a classifier.

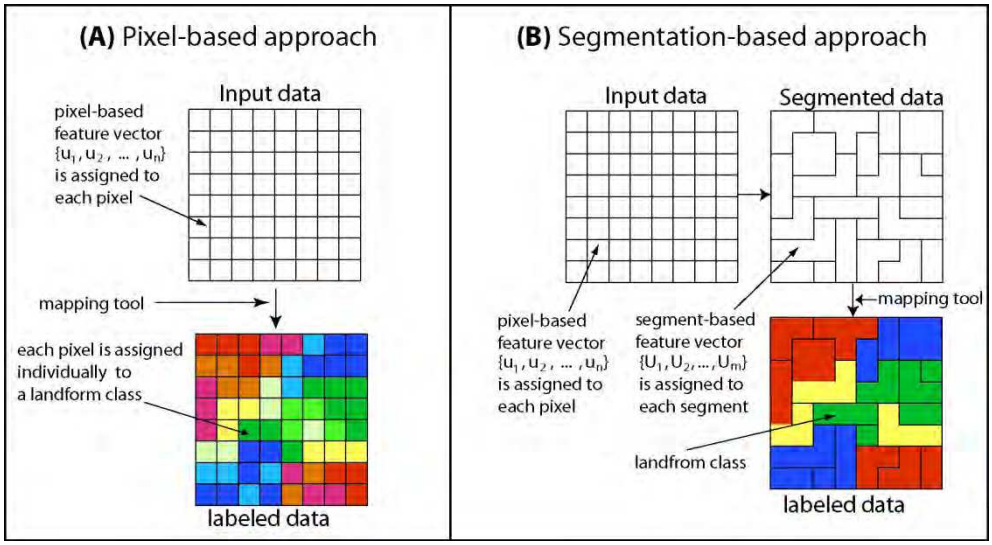

Fig. 1. Two different approaches to assigning geomorphic labels to topographic data

3. Unsupervised learning for exploratory mapping

Development of tools for exploratory geomorphic mapping of Martian surfaces is motivated by a desire for taking stock of all potential landforms present in the site regardless of their semantic meaning. Exploratory mapping is best achieved by unsupervised learning that relies on clustering techniques to automatically discover natural clusters in data. We discuss generating exploratory maps of Martian geomorphology using both, pixel-based and segmentation-based method.

3.1 Pixel-based, unsupervised mapping

In our pixel-based application to exploratory mapping of Mars (Stepinski and Vilalta, 2005), an unsupervised learning algorithm groups pixels that are similar in the space of geomorphic features. The choice of features is dictated by the type of surface to be mapped. A large portion of Martian surface consists of cratered plateau; planetary geomorphologists are interested in mapping various parts of craters, non-crater ridges, linear landforms known as channels, and various types of plateau. This interest dictates the choice of features that are best to discriminate between potential landforms of interest.

In the first tool that we have developed (Stepinski and Vilalta, 2005) the mapping is based on six features (terrain attributes). The first feature, u_1, is the elevation itself. The second feature, u_2, is a "flooding adjustment"; In order to calculate u_2 we first artificially modify the original elevation using the so-called "flooding" algorithm (O'Callaghan and Mark, 1984). It identifies all enclosed topographic depressions and raises their elevation to the level of the lowest pour point around their edge thus producing a "flooded" elevation field. The flooding adjustment (u_2) is the difference between flooded and original elevations; it has non-zero values only for pixels located inside topographic depressions (craters). The third feature, u_3, is the steepest slope between a focus pixel and the eight of its nearest neighboring pixels calculated using the original elevation field. The fourth feature, u_4, is the steepest slope calculated using the flooded elevation field. The fifth feature, u_5, is a contributing area. The contributing area is the total number of pixels "draining" through a focus pixel; the term draining is used here as a metaphor for connectivity between different pixels in a landscape. A pixel counts toward the contributing area of a focus pixel if there is a chain of steepest slope directions linking it to the focus pixel. Small values of u_5 flag pixels located on topographic peaks, ridges, and divides. Large values of u_5 flag channels. Finally, the sixth feature, u_6, is the contributing area based on the flooded elevation field.

The set of six features $\{ u_1, u_2, u_3, u_4, u_5, u_6 \}$ is calculated for all pixels in a site. The basic object of clustering is a pixel in the DEM that carries a vector of six features. Two pixels are similar if their feature vectors are close in the sense of Euclidean metric. A clustering algorithm applied for all pixels produces as output a set of k classes, $C_k = \{ c_1, c_2, \ldots, c_k \}$ where each class c_i contains a list of pixels that are similar to each other. The set of classes is mutually exclusive and exhaustive. The map is generated by assigning each pixel a color corresponding to its class. In our first implementation (Stepinski and Vilalta, 2005) of our pixel-based exploratory mapping tool we cluster the DEM using probabilistic clustering algorithm that follows the Expectation Maximization (EM) technique (McLachlan and Krishan, 1997). It groups vectors into classes by modeling each class through a probability density function. Each vector in the dataset has a probability of class membership and is assigned to the class with highest posterior probability. The number of classes is calculated using cross-validation (Cheesman and Stutz, 1996). Because a typical Martian DEM of interest contains a large amount of data, a direct clustering via the EM technique is computationally expensive. To alleviate this problem we sample the DEM to create a smaller, initial dataset of pixels. This initial dataset is clustered into C_k using the EM technique. The remaining pixels are classified into C_k using a decision tree learning algorithm (Quinlan, 1993) constructed on the basis of the initial dataset.

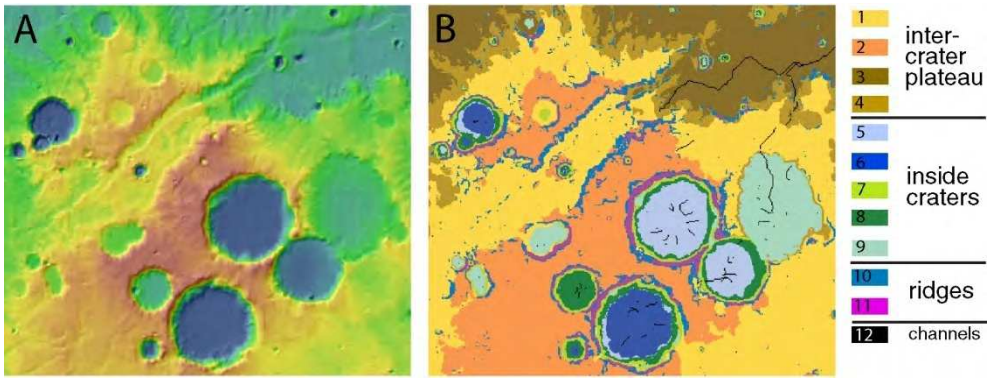

Fig. 2. (A) Topography of Tisia Valles site on Mars. (B) The 12-class geomorphic map created by a pixel-based tool using a probabilistic clustering algorithm.

In order to demonstrate the utility of our tool for producing an exploratory geomorphic map of landscape on Mars we have applied it to a test site called Tisia Valles. The six-feature vector was calculated for each of the site's 163,240 pixels. Because different features have different physical meaning and different range of values, we have normalized all features so that their values are in the range (0, 1). This normalization assures that every feature contributes with equal weight to the "distance" between different pixels. The 40,000 pixels were randomly chosen to create an initial dataset. We have assured uniform sampling in order to obtain an unbiased representation of all, even rare landscape features. Our clustering algorithm has grouped these 40,000 pixels into 12 separable and exhaustive clusters. The remaining pixels were classified into those 12 clusters using a decision tree algorithm.

Fig. 2A. shows the topography of the test site; red-to-blue gradient indicates high-to-low elevation. Fig. 2B. shows a geomorphic map generated by our tool; different landform classes (clusters of similar feature vectors) are shown using different colors. The semantic interpretation of these classes requires expert judgment; an analyst needs to review statistics of feature vectors values in each class and spatial distribution of classes with respect to each other. A simplified result of such interpretation is given in the legend of Fig. 2. An analyst divided the 12 classes into 4 different groups pertaining to plateau, craters, ridges, and channels, respectively. Some groups, for example the plateau group, may include several landforms classes. An expert grouped the four classes (labeled 1, 2, 3, and 4) into the plateau group because they are identical from a geomorphic point of view, just located at different elevations. This example illustrates a "problem" with mapping based on the principle of unsupervised learning - a reasonable cluster derived under a proximity measure may not constitute a "novel" landform as perceived by an analyst. Nevertheless, in face of lack of any previous knowledge about the site's landscape, unsupervised learning delivers valuable, first draft information.

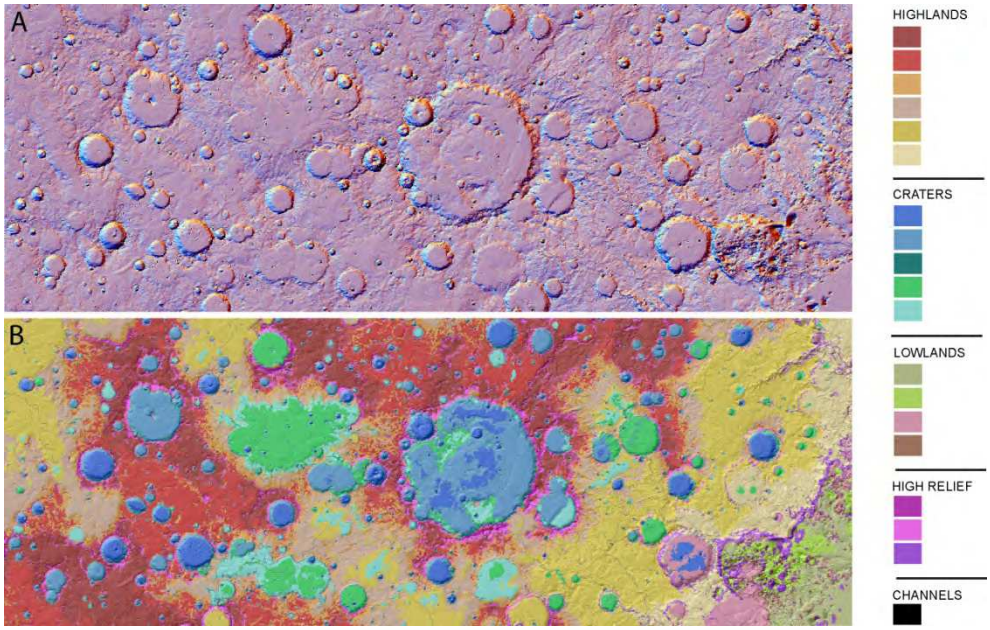

Fig. 3. (A) Topography of Terra Cimmeria site on Mars. (B) The 20-class geomorphic map created by a pixel-based tool using a SOM-Ward clustering algorithm.

Our tool for exploratory mapping using pixel-based unsupervised learning technique was further modified (Bue and Stepinski, 2006) to achieve greater computational efficiency necessary for mapping larger sites. The clustering efficiency was significantly improved by using a two-level clustering procedure (Vesanto and Alhoniemi, 2000) consisting of a self-organizing map (SOM) (Kohonen, 1995) and the Ward hierarchical clustering (Ward, 1963). The SOM is a neural network technique that groups similar vectors into nearby points on a 2-D grid composed of nodes. Through an unsupervised, iterative procedure, a large set of feature vectors is mapped onto the much smaller number of SOM's nodes in such a way that similar feature vectors are associated with the same note or neighboring nodes. Because the number of nodes is much smaller than the number of vectors, many similar vectors are mapped onto a single node. The bundle of feature vectors associated with a given SOM node is typified by a single representative vector referred to as a codebook vector. The final clustering of feature vectors into an assigned number of k clusters (landform classes) is achieved by the Ward's minimum variance grouping algorithm (Ward, 1963) applied to the set of codebook vectors.

Fig. 3A. depicts a topography of a 5,303,888 pixels-large site on Mars referred to as Terra Cimmeria. We used 30 x 30 rectangular SOM grid to perform a first step of clustering the feature vectors associated with those pixels into 900 codebook vectors. In the final step the codebook vectors were clustered into 20 landform classes shown on Fig. 3B using different colors. As in the previous example, the semantic interpretation of the classes requires expert judgment; a simplified result of such interpretation is given in the legend of Fig. 3.

3.2 Segmentation-based, unsupervised mapping

Our second tool for exploratory mapping via unsupervised learning combines aspects of pixel-based and segment-based mapping approaches (Stepinski and Bagaria, 2009). We constructed a two-stage algorithm consisting of a pixel-based *base* classifier and a segment-based *meta* classifier. A base classifier is applied to multiple pixels in a neighborhood of a focus pixel resulting in an ensemble of landform type predictions. A meta classifier is an unsupervised segmentation/classification algorithm that combines these predictions and outputs a segment-based map of emergent landform regions or classes. This tool is designed for exploratory mapping of very large regions using small number of original features.

In order to increase a computational efficiency of our tool we utilize a rule-based classifier as our base classifier. The rule-based classifier uses empirical knowledge to construct a decision tree; submitting a set of terrain features to a trunk of the tree results in a landform type label at the leave of the tree. The nested means technique (Iwahashi and Pike, 2007) is used to construct a decision tree because it outputs landform types whose meanings do not correspond directly to named terrestrial formations, thus, they won't lose their relevance in application to non-terrestrial surfaces. Our rule-based classifier uses only three original terrain features (slope gradient, surface texture, and local convexity) to label each pixel into one of 16 statistically predefined landform types.

The segment-based meta classifier uses a set of secondary features designed to capture contextual information around a given pixel. The secondary features are calculated from the labels (1 to 16) returned by the base classifier; they are combined into a pixel-attached feature-vector which describes, in a generalized manner, surface character in the neighborhood of this pixel. We calculate 19 secondary features. The first 16 features are normalized frequencies of labels outputted by the base classifier contained within a N x N pixels square window centered on the focus pixel. The value of N controls the level of generalization from landform types to landform classes. Two windows may have similar frequencies but different spatial distributions of the labels. The last three secondary features measure pattern of landform types in a neighborhood and are based on a modification of Multi-Scale Local Binary Pattern (LBP) concept (Ojala et al, 2002). The 19-dimensional vector of secondary features is used to generate a final segmentation-based map.

We use the Recursive Hierarchical Segmentation (RHSEG) algorithm (Tilton, 2000) that *simultaneously* segments the DEM on the basis of secondary features and cluster the segments into landform classes. The RHSEG is an iterative algorithm that produces hierarchies of both, segmentation levels, and clustering levels. Starting from individual pixels as regions seeds, the algorithm alternates between merging similar adjacent regions into larger regions (segmentation) and merging labels of non-adjacent similar regions (clustering). Both steps utilize similarity criteria based on statistics of secondary features of pixels constituting the segments. As this two-step process is iteratively repeated, it produces a natural hierarchy of both, spatial segmentations and clusters of features. Stopping RHSEG at a given iteration level yields a map of a certain geographical and feature-space resolution.

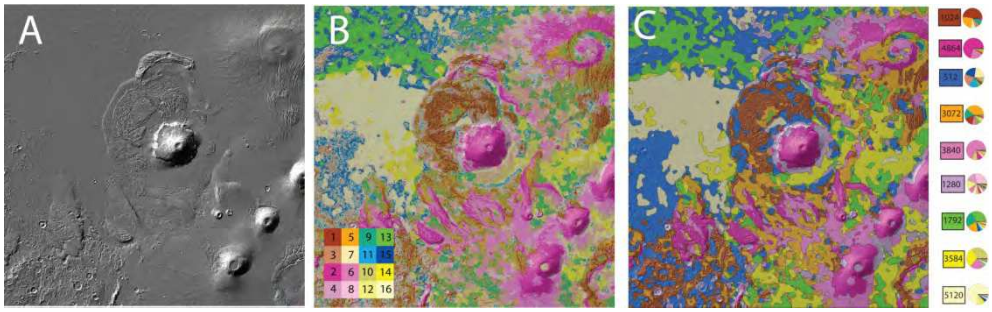

Fig. 4. (A) Topography of Tharsis region on Mars. (B) The results of rule-based classification. (C) The 9-class geomorphic map created by the meta classifier.

In order to demonstrate the utility of our segmentation-based exploratory mapping tool we have applied it to the Tharsis region on planet Mars. The Tharsis region on Mars (Fig. 4A) is an enormous volcanic plateau containing a number of large volcanoes including Olympus Mons – the largest volcano in the solar system. We used a 1024 x 1024 pixels DEM of Tharsis region with the resolution of 4 km/pixel. The base classifier labeled each of 1,048, 576 pixels with one of 16 labels resulting in a pixel-based map as shown on Fig. 4B. The legend to Fig. 4B is organized in a square array (see insert): the top row groups terrain types (1, 5, 9, 13) representing rough, convex landscapes; the second row groups terrain types (3, 7, 11, 15) standing for rough, concave landscape; the third row groups terrain types (2, 6, 10, 14) representing smooth, convex landscapes; the last row groups terrain types (4, 8, 12, 16) corresponding to smooth, concave landscapes. In each row progressively larger values of labels indicate gentler landscape. The secondary features are calculated using N=11 pixels moving window. We set the parameters of RHSEG algorithm so it starts saving the segmentation results when the feature-vectors are already clustered into 20 landform classes. This most-detailed of all retained partitioning is referred to as level 0 segmentation. Subsequent, progressively coarser segmentations are referred to as level 1 to 19, respectively. Fig. 4C shows the level 11 segmentation consisting of 2382 segments grouped into 9 landform classes. The legend to Fig. 4C shows a color and a numerical label assigned to each landform class.

Notwithstanding superficial visual similarity (this similarity decreases rapidly when the close-ups of the two maps are examined) between the map generated by the base classifier (Fig. 4B) and the map generated by the meta classifier (Fig. 4C), the map generated by RHSEG algorithm has a higher utility because it partitions a site in a fashion similar to what an analyst would do manually – into fewer larger, more heterogeneous areas corresponding to more broadly defined landform classes. The small pie diagrams next to label annotations in a legend to Fig. 4C indicate a "composition" of each landform class in terms of types outputted by the base classifier. Different classes are characterized by different degrees of terrain type inhomogeneity reflecting the reality of natural landscape. The visual esthetics of the map shown on Fig. 4C resembles manually drawn geologic maps, however, direct, formal comparison of our map with a manually drawn geologic map is difficult because analysts use not only objective criteria (such as, surface morphology) but also subjective criteria (such as, nomenclature, history of previous investigations, etc.). Nevertheless, our

map shown on Fig. 4C shows a rough correspondence to a manually drawn geologic map of the Tharsis region (Scott and Tanaka, 1986).

4. Supervised learning for exploitation mapping

In many cases planetary scientists know a priori what landform classes they want to map in a given site. Automation of such exploitation mapping should not be based on the principle of unsupervised learning that offers no control over the character of outcome classes; instead, it should be based on the principle of supervised learning where classes are set a priori. Recognizing that automatically generated maps must conform to expectations of the planetary science domain, our efforts to automate the process of exploitation mapping focus on the concatenation of a segmentation-based technique with supervised learning (Stepinski et al, 2006; Stepinski et al, 2007; Ghosh et al, 2009). The idea of segmentation-based classification follows from the realization that pixels are not the best fundamental units of visual or topographic scenes, and it is more natural and efficient to work with more perceptually meaningful entities obtained from low-level grouping processes. Such entities are referred to as superpixels (Mori, 2005) in the computer vision community and as segments (Benz et al, 2004) in the remote sensing community. The diagram in Fig. 1B illustrates the concept of segmentation-based classification of landform classes. The segmentation-based classification technique has many desired properties: a) segments are perceptually meaningful, b) they are computationally efficient, c) their geometric and statistical properties provide additional information that can be incorporated into classification, d) because the technique results in oversegmentation of the site, most structures in the site are conserved and there is little loss of information over using individual pixels.

We have developed a family of tools for automating exploitation mapping of planetary surfaces; each tool utilizes a specific combination of segmentation and classification algorithms. We employ two different segmentation algorithms. The dividing algorithm splits the landscape on the basis of abrupt discontinuities in pixel-based feature vectors. The agglomerative algorithm initially treats each pixel as an individual segment; these initial segments are combined into larger segments as long as a user-defined criterion for the uniformity of constituent pixel-based feature vectors holds. Both algorithms use the same pixel-based feature vectors. We also employ three different learning algorithms for segment classification and to generate maps of landforms. Thus, altogether, we have evaluated six different tools, corresponding to six different segmentation/classification combinations.

4.1 Segmentation methods

The segmentation procedure subdivides the landscape into mutually exclusive and exhaustive segments containing pixels having approximately uniform pixel-based feature vectors. These segments constitute topographic objects, which, subsequently, are classified into landforms classes. Raster segmentation has been the subject of intense study in the domain of image analysis, however, requirements for an effective segmentation for the purpose of mapping are different from those encountered in the field of computer vision. In particular, for the purpose of mapping-by-classification, it is desirable to oversegment the site. Having small segments eliminates the danger of a particularly large segment being misclassified, which would avoid producing a grossly incorrect map. Moreover, having

approximately equal-sized segments assures that statistics of pixel-based features are calculated from comparable ensembles of member pixels.

Our dividing segmentation algorithm (Stepinski et al, 2006) uses the watershed transform (Beucher, 1992) applied to a gray-scale image that encapsulates gradients of pixel-based feature vectors. This image is calculated using a computationally simple homogeneity measure H (Jing et al, 2003). A pixel located in a region that is homogeneous with respect to pixel-based features has a small value of H. On the other hand, a pixel located in a region which is inhomogeneous with respect to features has a large value of H. A raster constructed by calculating the values of H for all pixels in the landscape can be interpreted as a gray-scale image and is referred to as the H-image. White areas in H-image represent boundaries of homogeneous regions, whereas the dark areas represent the actual regions. The watershed transform of H results in (over) segmentation of the H-image (and thus the landscape).

Our agglomerative segmentation algorithm (Stepinski et al, 2007) uses a contiguity-enhanced variant of the standard K-means clustering algorithm, which uses – in addition to terrain attributes – spatial coordinates of pixels as features. The additional spatial features control the size of the segments while providing the resultant segments with very desirable geometric properties. For example, in areas where terrain features are approximately uniform, the local gradient of the total feature vector is dominated by changes in spatial coordinates leading to the formation of round-shaped segments. On the other hand, in areas where change in the total feature vector is dominated by change in terrain attributes, segments tend to exhibit an elongated shape in direction perpendicular to the gradient of the terrain-only sub-vector. These properties constitute additional knowledge that could be exploited by the classification module. The actual segmentation invokes a simple K-means algorithm applied to spatially-enriched, pixel-based feature vectors. The size of the segments is controlled by the value of K (which needs to be large to achieve over-segmentation).

4.2 Application of segmentation methods

In order to demonstrate the working of segmentation algorithms in practice we applied them to the Tisia Valles site on Mars (see Section 3.1). The site is segmented on the basis of three pixel-based terrain features $\{ u_1, u_2, u_3 \}$ using both, watershed and K-means, algorithms. Note that the featured used here are different from those we choose for exploratory mapping (see Section 3.1); they are: u_1=slope, u_2=curvature, and u_3=flooding adjustment. The watershed algorithm produced 7708 segments with sizes ranging from 1 to 267 pixels, whereas K-means algorithm (with the value of K = 5000) produced 6593 single-connected segments having sizes ranging from 4 to 117 pixels. Note that the K-means algorithm yields more than K segments because the resulting K clusters do not correspond to K single-connected spatial segments. In order to derive the segmentation we assign a unique segment identifier to each subset of a cluster corresponding to a single spatially connected region. The two algorithms yield segmentations having very different characters (see Fig. 5.). The watershed algorithm yielded a mosaic of segments that, by themselves, do not revel the landforms present in the site. On the other hand, the inclusion of spatial coordinates into the K-means algorithm resulted in segments that reflect the geometry of the landforms – one can notice the major landforms just from the segmentation image.

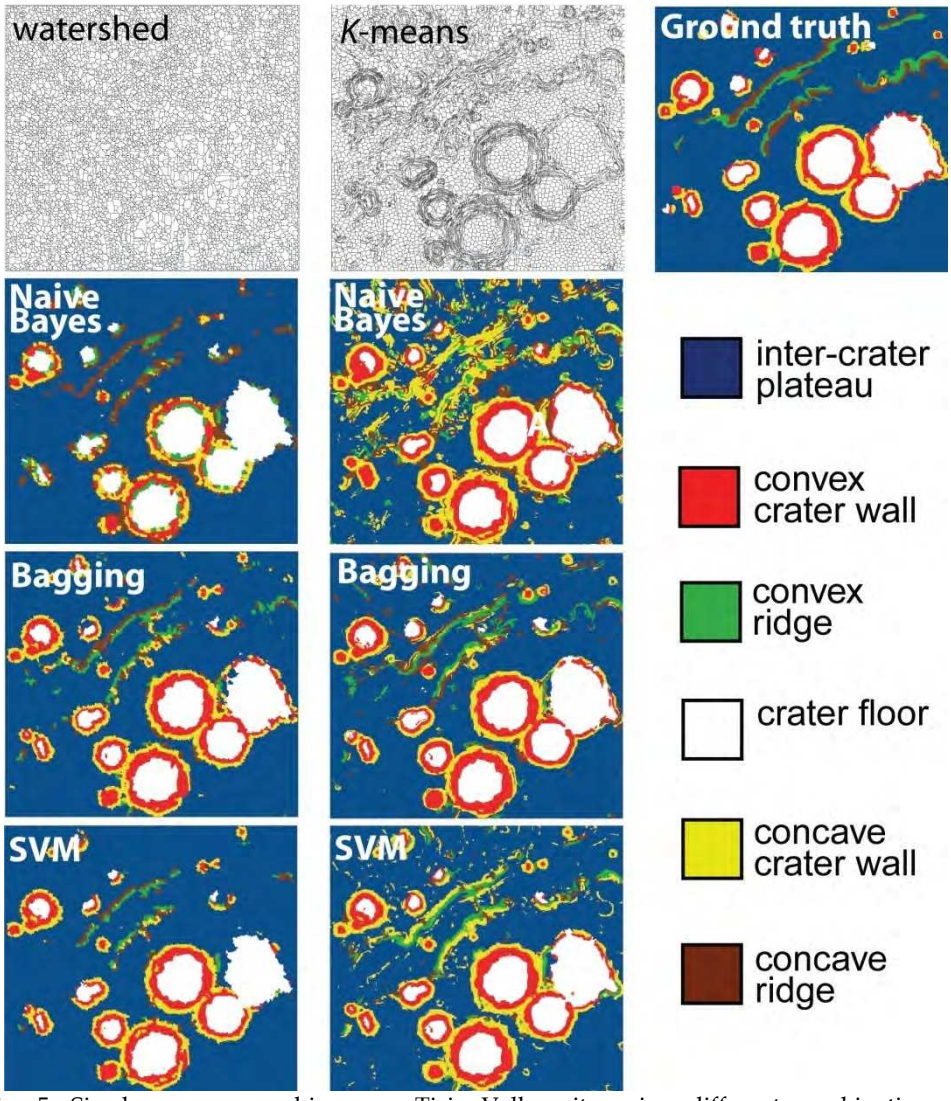

Fig. 5. Six-classes geomorphic maps Tisia Valles site using different combination of segmentation and classification algorithms.

4.3 Segment-based features

In the segmentation-based classification, pixel-based features used for segmentation are different from segment-based features used for classification. Each segment, regardless of an algorithm used to obtain the segmentation, is represented by a combination of physical and spatial segment-based features. Physical features are pixel-based features averaged over the

constituent pixels of the segments. Spatial features are obtained using each segment's shape measure and the neighborhood context measure. The shape measure is computed in terms of the Shape Complexity Index (SCI). The *SCI* is a measure of segment circularity. The closer the value of *SCI* is to 1.0, the more circular the object; on the other hand, thin ring-like shapes tend to have *SCI* values of 2.5 and higher. One of the challenges of the automatic classification of landforms is feature similarity of some landform classes that differ mostly by their spatial context. For instance, segments making up craters' walls and segments constituting ridges not associated with craters may have similar values of slope, curvature, but are located in different spatial contexts. In our segmentation-based tool we take into consideration spatial context by means of neighborhood context measures. Ideally, we would like to know classes of segment's neighbors to establish its spatial context, but such information is not available prior to classification. However, we can categorize the unlabeled segments into low, medium, and high categories based on statistics of the values of their physical features. Such categorization is used to calculate the neighborhood property of each segment using a nine-dimensional vector { a_h^s, a_m^s, a_l^s, a_h^c, a_m^c, a_l^c, a_h^f, a_m^f, a_l^f }, where a_j^i j=h, m, l and i= s (slope), c(curvature), f(flooding adjustment) is a percentage of the focus segment boundary with neighbors belonging to category high (h), medium (m), or low (l), respectively. Thus, a segment-based feature vector has 13 components, three physical features, the value of SCI, and 9 values of a_j^i.

4.4 Classification and mapping

We applied three different learning algorithms for segment classification and to generate geomorphic maps. First, the simple Naive Bayes algorithm provides a baseline for comparison with other classifiers. Second, the Support Vector Machines (SVM) algorithm that works by finding an optimal hyper-plane in a (transformed) feature space (Boser et al, 1992). The optimal hyper-plane maximizes the separation between classes. SVM exploits local data patterns and has been found to be effective in spatial data mining applications (Sharifzadeh et al, 2003). Third, bagging ensemble learning algorithm (Breiman, 1996) generates multiple models by running a single learning algorithm multiple times over bootstrapped samples of the training set. The final class label is the result of voting over the contributing models (one from each bootstrap sample). Bagging is known to work well for complex datasets and is particularly attractive when the training set is noisy (Dietterich, 2000). We use a decision tree (C4.5) as the base learner in the bagging algorithm.

We applied these classifiers to segments generated by watershed and *K*-means generated divisions of the Tisia Valles site. We have chosen six landform classes for mapping: crater floors, convex crater walls, concave crater walls, convex ridges, concave ridges, and inter-crater plateau. The choice of these particular landform classes stems from our interest in the quantitative characterization of old, cratered Martian surface. The labeled (training) set of segments was generated by manually labeling 30% (by surface area) of the Tisia site into the aforementioned six classes. Fig. 5 offers a visual assessment of the maps generated by different combination of segmentation and classification algorithms. The "ground truth" map of Tisia (an extension of the training set to the entire site) was hand-labeled. It shows how a typical analyst would map the six landforms in this site; it does not really constitute a ground truth (in the strict meaning of the concept) because an analyst is likely to draw an idealized map that misses details and projects a human conceptualization of the entire landscape, even if it contradicts local measurements. Maps based on the watershed

segmentation have a "simple" look as they lack small-scale details, whereas maps based on the *K*-means segmentation look exhibit more small-scale details. On the basis of only a visual inspection one could conclude that maps stemming from watershed segmentation are "better" because they look more like the ground truth map. However, closer inspection of the generated maps shows that maps based on *K*-means segmentation correctly reflect some small-scale details that are absent from the watershed segmentation and the analyst-drawn map. The maps generated by Naive Bayes are inaccurate and inferior to maps generated by Bagging and SVM.

watershed based segmentation							
classifier	overall accuracy	plateau	floor	cvx. wall	con. wall	cvx. ridge	con ridge
NB	84.67	0.94/0.97	0.93/0.93	0.87/0.39	0.57/0.37	0.12/0.12	0.20/0.51
B	89.31	0.95/0.96	0.97/0.87	0.80/0.83	0.65/0.75	0.51/0.45	0.38/0.37
SVM	89.71	0.92/0.98	0.96/0.90	0.87/0.75	0.72/0.70	0.62/0.29	0.48/0.29
K-means based segmentation							
classifier	overall accuracy	plateau	floor	cvx. wall	con. wall	cvx. ridge	con ridge
NB	74.11	0.99/0.76	0.99/0.80	0.70/0.78	0.32/0.83	0.14/0.19	0.08/0.25
B	87.42	0.96/0.92	0.94/0.92	0.80/0.85	0.67/0.72	0.34/0.47	0.25/0.40
SVM	86.10	0.97/0.91	0.98/0.87	0.77/0.79	0.46/0.83	0.46/0.54	0.18/0.12

Table 1. Assessment of performance of different methods used to map the Tisia Valles site. The entries for individual landform are precision/recall. NB – Naïve Bayes, B – Bagging with C4.5, SVM – Support Vector machines.

Table 1 gives accuracy rates for maps of the Tisia site. Disregarding maps produced by the Naive Bayes algorithm, accuracy rates are above 86%. Note that maps based on the watershed segmentation have slightly higher rates than maps based on the *K*-means segmentation in line with their greater similarity to the analyst drawing. Precision and recall rates for six landform classes are also given in Table 1. Results show that inter-crater plateau, crater floor, and convex crater walls landforms are mapped with high accuracy. Concave crater walls are detected with less accuracy, and ridges are difficult to identify correctly. This is because local ridges look like crater walls, even though they are different landforms in the context of the entire landscape.

5. Summary and conclusion

Geomorphic auto-mapping of planetary surfaces is a challenging problem. Here we have described how machine learning techniques, such as clustering or classification, can be utilized to automate the process of geomorphic mapping for exploratory and exploitation purposes. Relatively coarse resolution of planetary topographic data limits the number of features that can be used in the learning process and makes planetary auto-mapping more challenging than terrestrial auto-mapping. With this caveat, the methods discussed here are also applicable to terrestrial surfaces.

The major challenge in exploratory (unsupervised learning) mapping is to generate a map that has an appearance and utility similar to maps already used by the geosciences community. This means that a clustering algorithm should be able to generalize from a simple similarity of feature vectors to a similarity of ensembles of feature vectors. In other

words, an algorithm should be able to generate classes of varying degree of homogeneity based on spatial considerations. This is what our algorithm described in section 3.1 has been designed to do. Future research will address better criteria for deciding which classes should be homogeneous and which should be more heterogeneous. Overall, our two-stage tool for exploratory mapping is expected to be adopted by the geosciences community because it is matured enough for immediate application in geologic mapping, quantitative comparative geomorphology, and landscape visualization.

The major challenge in exploitation mapping (supervised learning) is the issue of spatial context. An analyst can map landforms having very similar features as different classes depending on broader spatial context. Thus, spatial context must be incorporated into the mapping algorithm in order to generate maps similar to those that are manually drawn. We have demonstrated that a choice of a particular segmentation method and a particular (capable) classification algorithm results in somewhat different maps, but, in general, all generated maps were acceptable. Indeed, regardless of the segmentation/classification combination, most misclassifications were the results of confusion due to spatial context. Our simple method of taking some account of spatial context proved insufficient to prevent misclassifications between elements of crater walls and elements of ridges. Future work needs to investigate more robust approach, such as, for example, Markov Random Fields, to incorporate spatial context information (Besag, 1986) into the learning algorithm.

Acknowledgements: This work was supported by the National Science Foundation under Grants IIS-0812271 and by NASA under grant NNG06GE57G. A portion of this research was conducted at the Lunar and Planetary Institute, which is operated by the USRA under contract CAN-NCC5-679 with NASA. This is LPI Contribution No. 1492.

6. References

Adediran, A.O.; Parcharidis, I.; Poscolieri, M. & Pavlopoulos, K. (2004). Computer-assisted discrimination of morphological units on north-central Crete (Greece) by applying multivariate statistics to local relief gradients. *Geomorphology*, Vol.58, pp.357-370.

Araki, H.; Tazawa, S.; Noda. H.; Ishihara, Y.; Goossens, S.; Kawano, N.; Sasaki, S.; Kamiya, I,.; Otake, H.; Oberst, J. & Shum, C.K. (2009). The lunar global topography by the laser altimeter (LALT) onboard Kaguya (Selene): Results from the one year observations. *Proceedings of 40th Lunar and Planetary Science Conference*, #1432.

Besag, J. (1986). On the statisical analysis of dirty pictures. *Journal of the Royal Statistical Society B*, Vol.48, pp.259–302.

Benz, U.; Hoffmann, P.; Willhauck, G.; Lingenfelder, I. & Heynen, M. (2004). Multi-Resolution, Object-Oriented Fuzzy Analysis of Remote Sensing Data for GIS-Ready Information, *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol.58, pp.239-258.

Beucher, S. (2003). The watershed transformation applied to image segmentation. *Scanning Microscopy International*, Vol.6, pp.299–314.

Boser, B. E.; Guyon, I. & Vapnik, V. (1992). A training algorithm for optimal margin classifiers. *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pp.144–152.

Brown, D. G.; Lusch, D. P. & Duda, K. A. (1998). Supervised classification of types of glaciated landscapes using digital elevation data. *Geomorphology*, Vol.21, pp.233-250.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, Vol.24 No.2, pp.123–140.

Bue, B. D. & Stepinski, T. F. (2006). Automated classification of landforms on Mars. *Computers & Geosciences*, Vol. 32 No.5, pp. 604-614.

Burrough, P.A.; van Gaans, P.F.M. & MacMillan, R.A. (2000). High-Resolution landform classification using fuzzy k-means. *Fuzzy Sets and Systems*, Vol.113, pp.37–52.

Cheesman, P. & Stutz, J. (1996). Bayesian Classification (AutoClass) Theory and Practice. In: *Advances in Knowledge Discovery and Data Mining*, U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy (Eds.), pp.153-180, MIT Press.

Dietterich, T.G. (200). Ensemble methods in machine learning. In: *Lecture Notes in Computer Science # 1857*, pp.1–15.

Dragut, L. & Blaschke, T. (2006) Automated classification of landform elements using object-based image analysis. *Geomorphology*, Vol.81, pp.330–344.

Evans, I.S. (1998), What do terrain statistics really mean?, In: *Landform monitoring, modelling and analysis*, Lane, S.N., Richards, K.S., and Chandler, J.H., (Eds.), pp. 119-138, J.Wiley, Chichester.

Ehsani, A.H. & Quiel, F. (2008). Geomorphometric feature analysis using morphometric parameterization and artificial neural networks. *Geomorphology*, Vol.99, pp.1–12.

Gallant, A.L.; Brown, D.D. & Hoffer, R. M. (2005). Automated mapping of Hammond's landforms. *IEEE Geoscience and Remote Sensing Letters*, Vol.2 No.4, pp.384–288.

Ghosh, S.; Stepinski, T. F. & Vilalta, R. (2008). Automatic Annotation of Planetary Surfaces with Geomorphic Labels. *IEEE Transactions on Geoscience and Remote Sensing*, submitted.

Hengl, T. & Rossiter, D.G. (2003). Supervised landform classification to enhance and replace photointerpretation in semi-detailed soil survey. *Soil Science Society of America Journal*, Vol.67, pp.1810-1822.

Irvin, B.J.; Ventura, S.J. & Slater, B.K. (1997). Fuzzy and isodata classification of landform elements from digital terrain data in Pleasant Valley, Wisconsin. *Geoderma*, Vol. 77, pp. 137-154.

Iwahashi, J. & Pike, R. J. (2007). Automated classifications of topography from DEMs by an unsupervised nested-means algorithm and a three-part geometric signature. *Geomorphology*, Vol. 86, pp. 409-440.

Jing, F.; Li, M.; Zhang, H. & Zhang, B. (2003). Unsupervised image segmentation using local homogeneity analysis. *Proceedings of 2003 International Symposium on Circuits and Systems*, pp.II-456-II459.

Kohonen, T. (1995). *Self-organizing Maps*. Springer, Berlin.

Krishna, B.G.; , Amitabh; Singh, S.; Srivastava, P.K. & Kiran Kumar, A.S. (2009). Digital elevation models of the lunar surface from Chandrayan-1 terrain mapping camera (TMC) imagery – initial results. *Proceedings of 40th Lunar and Planetary Science Conference*, #1694.

McLachlan, G. & Krishnan, T. (1997) *The EM Algorithm and Extensions*. John Wilay and Sons, New York, NY.

Mori, G. (2005). Guiding Model Search Using Segmentation. *Proceedings of the Tenth IEEE International Conference on Computer Vision*, pp.1417 - 1423.

O'Callaghan, J.F. & Mark, D.M. (1984). The extraction of drainage networks from digital elevation data. *Computer Vision, Graphics and Image Processing,* Vol.28, pp.328–344.

Ojala, T.; Pietikainen, M. & Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 24 No.7, pp.971–987.

Prima, O.; Echigo, A.; Yokoyama, R. & Yoshida, T. (2006). Supervised landform classification of northeast Honshu from DEM derived thematic maps. *Geomorphology,* Vol.78, pp.373–386.

Quinlan, J.R. (1993). *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Francisco.

Scott, D.H. & Tanaka, K.L. (1986). Geologic map of the western equatorial regions of Mars, *US Geol. Surv. Inv. Series Map*, I-1802 A.

Sharifzadeh, M.; Shahabi, C. & Knoblock, C. (2003). Learning approximate thematic maps from labeled geospatial data. *Proceedings of International Workshop on Next Generation Geospatial Information*, Cambridge (Boston), Massachusetts, USA.

Smith, D.; Neumann, G.; Arvidson, R.; Guinness, E. & Slavney, S. (2003). Mars global surveyor laser altimeter mission experiment gridded data record. *NASA Planetary Data System*, (MGS-M-MOLA-MEGDR-L3-V1.0), 2003.

Stepinski, T.F. & Vilalta, R. (2005). Digital topography models for Martian surfaces. *IEEE Geoscience and Remote Sensing Letters*, Vol. 2 No.3, pp. 260–264.

Stepinski, T.F.; Ghosh, S. & Vilalta, R. (2006). Automatic recognition of landforms on Mars using terrain segmentation and classification, *Proceedings of Ninth International Conference on Discovery Science*, Lecture Notes in Computer Science # 4265, pp. 255–266.

Stepinski, T.F.; Ghosh, S. & Vilalta, R. (2007). Machine learning for automatic mapping of planetary surfaces, *Proceedings of The Nineteenth Innovative Applications of Artificial Intelligence Conference*, AAI Press.

Stepinski, T.F. & Bagaria, C. (2009). Segmentation-based Unsupervised Terrain Classification for Generation of Physiographic Maps, *IEEE Geoscience and Remote Sensing Letters*, in press.

Tanaka, K.L. (1994). The Venus Geologic Mappers' Handbook, *U.S. Geol. Surv. Open File Rep.* pp.99–438.

Tilton, J.C. (2000), Method for recursive hierarchical segmentation by region growing and spectral clustering with a natural convergence criterion. In: *Disclosure of Invention and New Technology*, NASA Case Number GSC 14,328-1.

Wilhelms, D.E. (1990). Geologic Mapping. In: *Planetary Mapping*, Greeley. R.; and Batson, R. (Ed.), pp.209–260, Cambridge Univ. Press, Cambridge, UK.

van Asselen, S. & Seijmonsbergen, A.C. (2006.) Expert-driven semi-automated geomorphological mapping for a mountainous area using a laser DTM. *Geomorphology*, Vol.78, pp. 309-320.

Vesanto, J.; & Alhoniemi, E. (2000). Clustering of the self-organizing map. *IEEE Transactions on Neural Networks* Vol.11 No.3, pp.586–600.

Ward Jr., J.H. (1963). Hierarchical grouping to optimize an objective function, Vol.58 No.301, pp.236–244.

Machine Learning

Edited by Yagang Zhang

ISBN 978-953-307-033-9
Hard cover, 438 pages
Publisher InTech
Published online 01, February, 2010
Published in print edition February, 2010

Machine learning techniques have the potential of alleviating the complexity of knowledge acquisition. This book presents today's state and development tendencies of machine learning. It is a multi-author book. Taking into account the large amount of knowledge about machine learning and practice presented in the book, it is divided into three major parts: Introduction, Machine Learning Theory and Applications. Part I focuses on the introduction to machine learning. The author also attempts to promote a new design of thinking machines and development philosophy. Considering the growing complexity and serious difficulties of information processing in machine learning, in Part II of the book, the theoretical foundations of machine learning are considered, and they mainly include self-organizing maps (SOMs), clustering, artificial neural networks, nonlinear control, fuzzy system and knowledge-based system (KBS). Part III contains selected applications of various machine learning approaches, from flight delays, network intrusion, immune system, ship design to CT and RNA target prediction. The book will be of interest to industrial engineers and scientists as well as academics who wish to pursue machine learning. The book is intended for both graduate and postgraduate students in fields such as computer science, cybernetics, system sciences, engineering, statistics, and social sciences, and as a reference for software professionals and practitioners.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Tomasz F. Stepinski and Ricardo Vilalta (2010). Machine Learning Tools for Geomorphic Mapping of Planetary Surfaces, Machine Learning, Yagang Zhang (Ed.), ISBN: 978-953-307-033-9, InTech, Available from: http://www.intechopen.com/books/machine-learning/machine-learning-tools-for-geomorphic-mapping-of-planetary-surfaces

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821