Identifying Volcanoes From Elevation Profile

Elizabeth A. Culbertson
Stanford University
eaculb@stanford.edu

***Abstract*—In this paper, we compare two approaches to feature selection and three different supervised learning algorithms in the classification of volcanoes from a region's elevation profile. The best approach found was a Support Vector Machine used on the raw elevation values from a region. Finding an optimal classifier for the existence of volcanoes is important as it can help geologists identify regions of interest in hard to access areas such as the ocean floor and Mars.**

I. Introduction

A. Motivation

The presence or lack of volcanoes is useful in determining the geographic history of a region. Especially for unexplored regions such as the surface of Mars or the ocean floor, identifying and studying volcanoes present in the region is crucial to develop a full understanding of planetary evolution. This project explores the possibility of using features derived from elevation data to identify the presence of volcanoes. A model trained on known United States volcanoes could be used to identify areas of interest for geologists or volcanologists in areas that are hard to systematically explore, but have easily attainable elevation data using remote sensing.

B. Related Works

Statisticians, geologists, and geomorphologists have been attempting to classify primary topographical features of different types of landforms for decades. In 2006, Prima et al. used elevation data and raster maps of regions in Japan to identify different landforms (mountains, volcanoes, alluvial fans) by extracting ten morphometric parameters from their data. Their classification scheme ended up being based primarily on standard deviation of slope and topographic openness. Their approach had an 88% accuracy rate overall; however, they reported having the highest error rate (22%) for volcanoes, as a large number of volcanoes were mistakenly classified as non-volcanic mountains [1]. Dymond et al. established an algorithm for splitting an image of a landform into "land components" (regions of uniform slope) for polygonization of a landscape using elevation alone [2]. This method was successful in deriving useful information directly from contours; however, the report, published in 1994, cites insufficient computing power as a barrier for future research on the project, with no follow-up readily available that may have been performed later. More recently, Rathnam et al. performed a very similar analysis to what this paper wants to achieve; their team attempted to identify the location of volcano hotspots using sattelite imagery of the Earth's surface [3]. Their approach used an artificial neural network and standard back propagation learning on color satellite images to develop a classification scheme. This approach had an 82% accuracy rate and specifically a false positive rate of 16%. Bohnenstiehl et al. took a very different approach to volcano classification and applied it to the sea floor; their closed-contour model operated by "selecting the lowest elevation contour with a quasi-elliptical shape that completely encloses a topographic high" [4]. This approach is similar to ours in that it attempted to make a classification based on elevation contours alone; however, the learning scheme is quite different. Finally, as an interesting template for what we hope our model may achieve, Head et al. (1992) analyzed data from the Magellan satellite to extract key features of volcanoes on Venus and make a rudimentary clustering classification based on size and shape [5]. Our project touches on aspect from all of these works; it attempts to classify regions into volcanoes or non-volcanoes, with hopes of applying the model to regions such as other planets and the ocean floor, and aims to do so using remotely sensed elevation data alone.

C. Project Goals

In this project, we aimed to answer two main questions: 1. What is the best way to choose features for this problem? And 2. What is the best supervised learning model for classifying volcanoes using elevation data?

II. Data Acquisition and Processing

Elevation data was acquired for 115 volcanoes and 119 non-volcano regions using the USGS Bulk Point Query Service (BPQS). The latitude and longitude of the center of each region were run through a script to generate a 21x21 grid of coordinates centered around the input point and spanning a width of .1. This grid was then submitted to the BPQS to fill in the elevation at each coordinate. This resulted in 234 total examples each consisting off 441 latitude-longitude-elevation triplets. Examples of the positive and negative examples can be seen in figures (1a) and (1b).

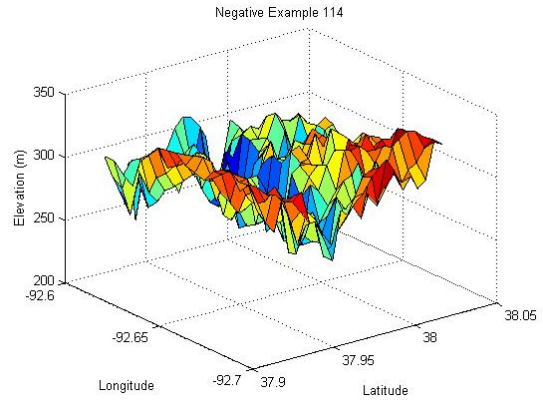
In general, qualitatively, the positive examples tended to look more sharply peaked and radially symmetric, while negative examples tended to be either relatively flat, or only be sloped along a particular direction. These observations motivated the feature selection.

III. Feature Selection

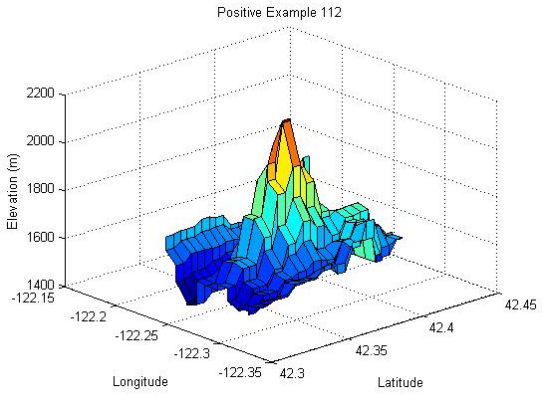
Two approaches were taken to extract features from the data: raw elevations and scores.

A. Raw Elevation Approach

The raw elevation approach was to let all 441 elevation measurements be the 441 features for each example, with location 1 on the grid being the first feature, location 2 the second, etc.



(a) Sample Negative Example (non-volcano)



(b) Sample Positive Example (volcano)

Fig. 1: Sample contour plots of both negative and positive training examples

B. "Scores" Approach

The other approach, in an attempt to quantify the key qualitative characteristics of a volcano, extracted two scores from the data to use as features:

1) Elevation Score: The goal of the elevation score is to get a sense of the slope of the region. In particular, it aims to distinguish between flat regions that may have a high rotational symmetry simply because they are flat and non-flat features that also have high rotational symmetry. The elevation score was calculated as follows:

$$score_{elevation} = elevation_{max} - elevation_{min} \quad (1)$$

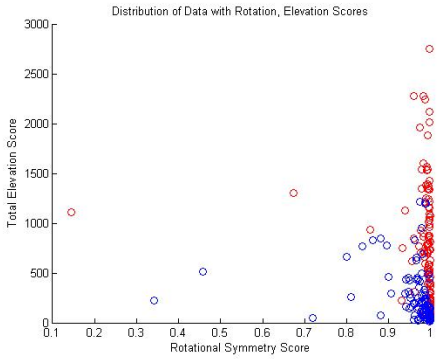
Other methods were tested as a way of determining the elevation score that were less effective. Namely, we attempted to use a method that compared the average elevation of the center ninth of the region to the average of the outer part of the region, hoping that this would mitigate the effects of outlying elevations; however, in the positive examples, this only exacerbated the problems that occurred when the region was not perfectly centered around the peak of the volcanoes.

2) Rotational Symmetry Score: The goal of the rotational symmetry score was to measure the roundness of the volcano profile. In particular, we saw the rotational symmetry as a good way to distinguish between ridge-shaped mountain ranges or cliffs and volcanoes, as such a distinction would not be able to be caught using the elevation difference alone. To obtain this score, four vectors were extracted from the training examples grid of coordinate-elevation pairs: the center column, the center row, the forward diagonal, and the backward diagonal. The dot product of each of these vectors with its mirror image were averaged and reported as the rotational symmetry score:

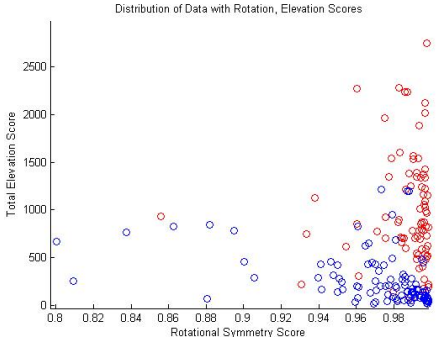
$$score_{symmetry} = \frac{1}{4}\sum_{i=1}^{4}(v_i \cdot flipped(v_i)) \quad (2)$$

Note that this score results in a higher number for a higher degree of rotational symmetry and normalizes all scores to a scale from 0 to 1. Other measurements were initially tested that assigned a higher score to regions with a lower degree of rotational symmetry: for example, one such version of this method was as such:

$$score_{symmetry} = \frac{1}{4}\sum_{i=1}^{4}\ln\frac{v_i}{flipped(v_i)} \quad (3)$$



(a) Datapoints plotted by scores. Red points are positive examples, blue points are negative examples.



(b) Same data points; zoomed in to the area of the graph with closely clustered points

Fig. 2: Data plotted in the elevation score-symmetry score plane

Unfortunately, this method resulted in scores that blew up for too-symmetric regions, and the normalized dot-product approach was chosen for the final model testing. A plot of

all training points in the rotational symmetry-elevation score plane is depicted in figure (2).

IV. Models

A. Logistic Regression

Our initial baseline model was logistic regression, i.e. we attempted to fit a vector such that the hypothesis took the form:

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}} \tag{4}$$

With x including an intercept term. We used MATLAB's glmfit function to do so. This resulted in a test set error of 55% for the raw elevation features and 23% for the contrived scores (see model comparison section for more details). Figure (3) depicts the 2-feature fit using logistic regression and figure (4) depicts the confusion matrices for the raw elevations and scores approaches, respectively. In particular, we noticed that the raw elevation method seemed to overwhelmingly predict a volcano classification, resulting in a large number of false positives.

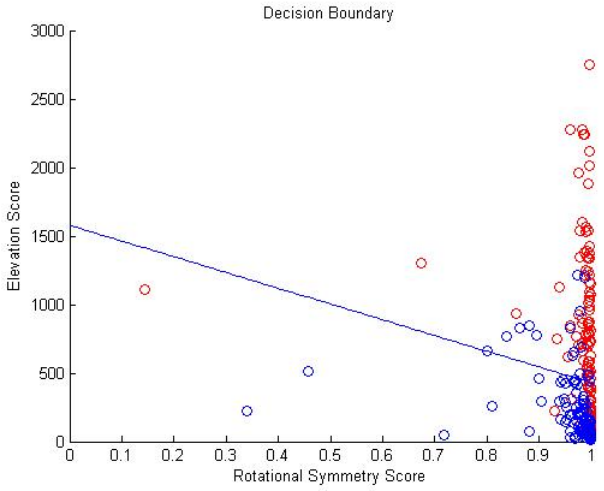


Fig. 3: Logistic Regression on Scores Approach

B. Gaussian Discriminant Analysis

The second approach used to model the data was Gaussian Discriminant Analysis, which models $p(x|y)$ as a multivariate normal distribution i.e.

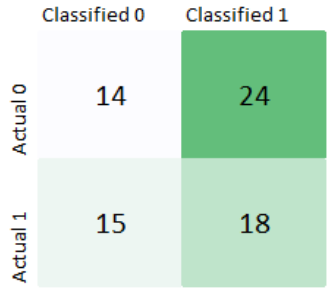
$$y \sim Ber(\phi) \tag{5}$$

$$p(x|y=0) \sim N(\mu_0, \Sigma) \tag{6}$$

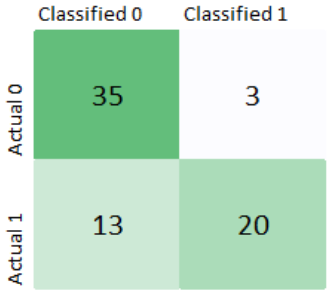
$$p(x|y=1) \sim N(\mu_1, \Sigma) \tag{7}$$

Where the parameters μ_0, μ_1, Σ, and ϕ are chosen by maximizing the likelihood of the parameters:

$$\phi = \frac{1}{m}\sum_{i=1}^{m} 1y^{(i)} = 1 \tag{8}$$



(a) Raw elevation



(b) Scores

Fig. 4: Confusion matrices for logistic regression

$$\mu_0 = \frac{\sum_{i=1}^{m} 1\{y^{(i)} = 0\}x^{(i)}}{\sum_{i=1}^{m} 1\{y^{(i)} = 0\}} \tag{9}$$

$$\mu_1 = \frac{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}x^{(i)}}{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}} \tag{10}$$

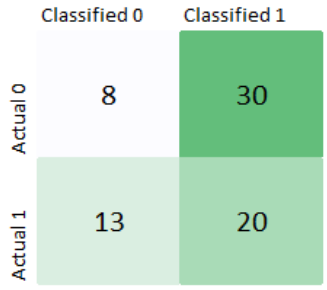
$$\Sigma = \frac{1}{m}\sum_{i=1}^{m}(x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T \tag{11}$$

We wrote the required code to accomplish this ourselves. This resulted in a test set error of 61% for the raw elevation features and 20% for the contrived scores (see model comparison section for more details). Again, this method tended towards many false positives (see figure (5) for confusion matrix.)

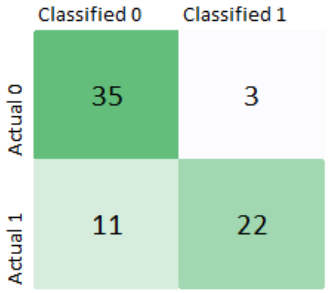
C. Support Vector Machine

The third classification method used was a Support Vector Machine (SVM) with an RBF kernel using MATLAB's fitcsvm function. With a feature space twice as large as the training set, this method resulted in nearly 90% error without regularization, so we chose to enforce L1 regularization. Thus, this method fit a weight vector w that solves the following optimization problem:

$$\min_{\gamma,w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i \tag{12}$$



(a) Raw elevation



(b) Scores

Fig. 5: Confusion matrices for GDA

$$\text{s.t. } y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i, i = 1 \ldots m \quad (13)$$

$$\xi_i \geq 0, i = 1 \ldots m \quad (14)$$

We empirically chose the box constraint C by comparing accuracy, precision, and recall measurements for varying values of C and choosing the value that maximized our performance matrix. Performance vs. box constraint plots are depicted in figure (6). Optimized box constraints for the SVM were found to be 1.75 and 0.75 for the 2-feature and 441-feature models, respectively. This resulted in a test set error of 15% for both the raw elevation features and the contrived scores (see model comparison section for more details). The SVM did not suffer the same false-positive problem as the other two models did for the raw elevation features (see figure (7) for confusion matrices).

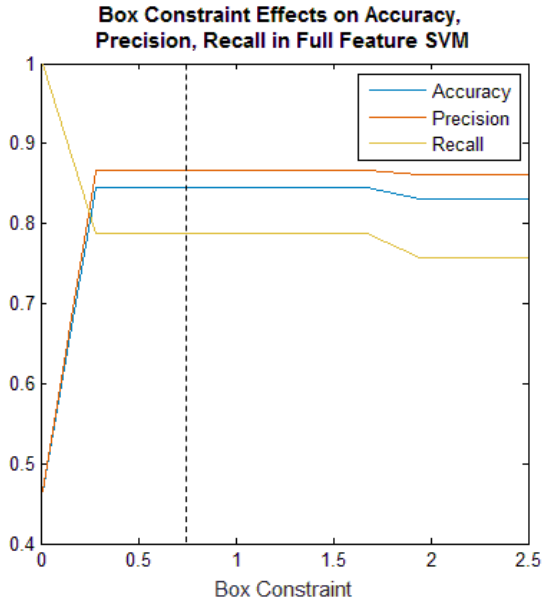
V. Results and Discussion

The tabulated accuracy, precision, and recall measurements for all feature-model combinations are in table (I). Figure (8) depicts this comparison graphically. All accuracy, precision, and recall results listed are testing error found using a 70%-30% train-test data split.

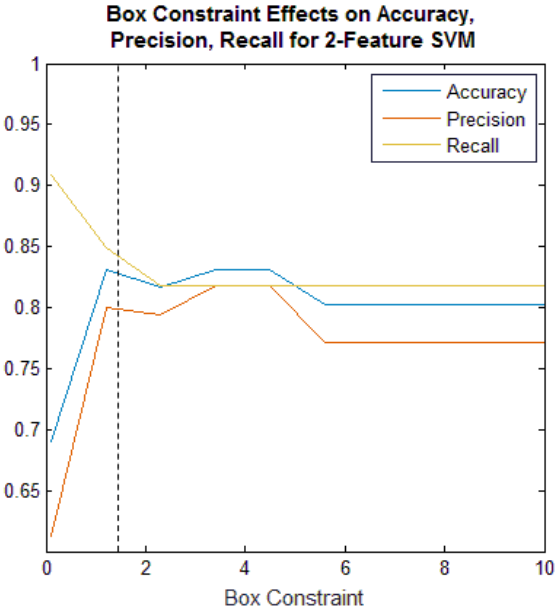
A. Feature Comparison

Using the raw elevation values as features rather than the contrived scores resulted in a marked tendency towards false positives. See figure () for a closer comparison between the two feature methods for logistic regression and GDA.

This result was not entirely surprising. We anticipated that using raw elevations, while technically retaining more



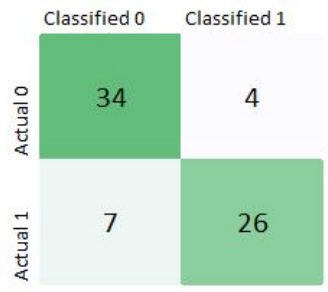
(a) Raw elevation



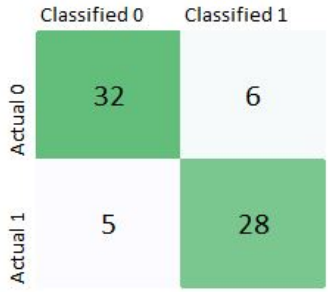
(b) Scores

Fig. 6: Regularization of the SVM: Effect of box constraint on performance metric (dotted line denotes the empirically optimal box constraint

information about the data, would not accurately capture the characteristics of a volcano without being mapped to a higher dimensional (not for the sake of increased dimensionality, but for the sake of more complexity) feature space. The scores reduced the amount of raw information fed to the model, but allowed the logistic regression and GDA models to learn on the interactions between different points elevations. Since the contour of a volcano is clearly related to the relationship between neighboring points above all else, it makes sense



(a) Raw elevation



(b) Scores

Fig. 7: Confusion matrices for SVM

TABLE I: Performance metrics of all feature-model pairs

Model	Feature Type	Accuracy	Precision	Recall
Logistic Regression	Raw Elevation	45%	43%	55%
Logistic Regression	Scores	77%	87%	61%
GDA	Raw Elevation	39%	40%	61%
GDA	Scores	80%	88%	67%
SVM with RBF kernel	Raw Elevation	85%	87%	79%
SVM with RBF kernel	Scores	85%	82%	85%

that the scores, which made claims about these relationships, outperformed the raw elevation approach. In addition, with a total data set of 234 and a training data set of 163, the raw elevations approach had an n on the order of and slightly greater than m while the scores approach had n an order of magnitude smaller than m. This means that the raw elevations approach was much more prone to overfitting than the scores. Because of these advantages of the scores using logistic regression and GDA, it makes sense that using the support vector machine, which both mapped to a higher dimensional feature space and reduced overfitting through regularization, equilibrated the performance of the two feature approaches. Interestingly, this also makes a comment on the choice of the contrived scores: their comparability to the raw elevation using the SVM indicates that they were a decent choice of extracted features in the first place.

B. Model Comparison

The SVM far outperformed logistic regression and GDA on testing accuracy, precision, and recall. This was in accordance with what we expected; the SVM potentially expanded the feature space of the scores approach while tackling the shortcomings of the raw elevations method.

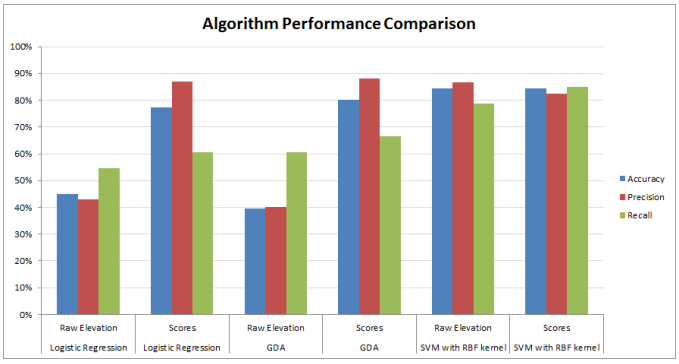


Fig. 8: Performance comparison of all feature-model combinations

VI. Conclusion and Future Work

A. Conclusion

The score approach to feature selection, perhaps due to its more nuanced interpretation of what constitutes a volcano and its lower likelihood of causing overfitting, outperformed the raw elevation approach using logistic regression and GDA. However, for the most successful model, the Support Vector Machine, the two approaches had similar performance, each obtaining a testing accuract of 85%.

B. Future Work

Given more time and a larger team, we would have liked to do a more formal feature selection/analysis of features. When fitting the various models to our data, we noticed that the parameter vectors were quite sparse, indicating that the dimensionality of the data could probably be reduced. As such, running PCA on the raw elevation data would be desirable. It would have also been nice to develop a few more contrived scores and do a more in-depth analysis on the effect of the different scores on the final model. It would also be interesting to see if our trained model could work on data from Mars and the ocean floor to identify areas of interest.

References: [1] Prima, Oky Dicky Ardiansyah et al. "Supervised Landform Classification of Northeast Honshu from DEM-Derived Thematic Maps." Geomorphology 78.34 (2006): 373386. ScienceDirect. Web. 14 Nov. 2015.

[2] Dymond, J. R., R. C. Derose, and G. R. Harmsworth. Automated Mapping of Land Components from Digital Elevation Data. Earth Surface Processes and Landforms 20.2 (1995): 131137. Wiley Online Library. Web. 14 Nov. 2015.

[3] VIA, STANDARD BACK PROPAGATION SBP ALGORITHM. "Identification of volcano hotspots by using standard back propagation (SBP) algorithm via satellite images." Journal of Theoretical and Applied Information Technology61.1 (2014).

[4] Bohnenstiehl, Delwayne R. et al. A Modified Basal Outlining Algorithm for Identifying Topographic Highs from Gridded Elevation Data, Part 1: Motivation and Methods. Comput. Geosci. 49 (2012): 308314. ACM Digital Library. Web. 12 Dec. 2015.

[5] Head, James W. et al. Venus Volcanism: Classification of Volcanic Features and Structures, Associations, and Global Distribution from Magellan Data. Journal of Geophysical Research: Planets 97.E8 (1992): 1315313197. Wiley Online Library. Web. 12 Dec. 2015.