Knowledge Mining and Soil Mapping
using Maximum Likelihood Classifier
with Gaussian Mixture Models

Jian Liu
December 7, 2005
Abstract:

(This study deals with data mining from soil survey maps and soil mapping with mined soil-landscape knowledge.)

Soil – landscape model describes the relationships between soil and its surrounding environments, based on which local soil can be predicated. Previous automated mapping approaches either 1) elicited such knowledge directly from a soil scientist and represented it with arbitrary curves (fuzzy system), or 2) used machine learning methods such as Artificial Neural Network (ANN) or decision trees to number-crunch existing maps to discover useful knowledge. However, the knowledge representation either was heuristic and lacked a sound justification (fuzzy system), or non-intuitive to comprehend (ANN), or crisp and non-flexible to fully capture the gradation of natural phenomenon (decision trees). This study proposes to represent soil-landscape knowledge with Gaussian Mixture Models (GMM), which well capture the physical nature of soil formation such as the interaction/correlation between environmental factors. The representations are “mined” from existing soil maps, and Maximum Likelihood (ML) classifier is used for mapping in a new area. A case study demonstrates the successful application of the proposed method. Performance of the algorithm is further discussed in the result section.
**Introduction**

Soil distribution is not random. Soil formations are dependent upon many environmental factors. For example, dry climate and wet climate yield different soils; slopes facing to the south might have different soils than slopes facing to the north; soils in the drainage way are usually different from those on the ridge tops. Traditional soil survey is based on such soil-environment relationships. Even though extensive field investigation is preformed during each survey, soil is largely mapped in a predicative way given the observed environments.

Environmental factors that are influential on soil formation are scale-dependent. At a regional scale, climate largely decides the type of soil group, however, at a small scale, topography and bedrock geology is usually sufficient to represent the surrounding environment that affects the forming of soils. This relationship has been generalized into a paradigm called soil-landscape model (Hudson, 1992), which states that the local soil can be predicated given the surrounding landscape conditions sufficiently described.

A few topographic indices (also called terrain indices) can be used to describe the topography / landscape conditions. The commonly used ones are: elevation, slope gradient, profile curvature, planform curvature and slope aspect. Elevation can be easily obtained in the form of a Digital Elevation Model (DEM). Others can be computed from DEM. Slope gradient is the first derivative along the steepest slope; profile curvature is the second derivative along the steepest slope; planform curvature is the second derivative perpendicular to the direction of profile curvature. Slope aspect tells the facing direction of a slope. In addition, bedrock geology is also a significant soil formation factor.
Given the introduction above, soil mapping is all about discovering soil-landscape models, i.e. the local soil as a function of a few environmental factors.

**Motivation**

Traditional soil survey is largely a combination of extensive field investigation and manual drawing, and many automated approaches have emerged in recent years to mitigate human labor involved and expedite the mapping speed. Zhu et al. (2001) proposed a fuzzy system based approach, where the soil-landscape knowledge was directly elicited from a soil scientist and represented as a set of independent curves in terms of each environmental factor used. By assuming independence of each dimension, it neglects the interaction/correlation between environmental factors. In addition, the shapes of the curves are arbitrary and hardly justifiable. Zhu (2000), Behrens (2005) and Scull (2005) presented case studies using artificial neural networks (ANN) to train a classifier and perform soil mapping. However, knowledge was captured in a black box in the form of a high dimensional matrix, which is hard to interpret and casts no insights into comprehensible soil-landscape relations. Decision trees were also tested (Bui 1999; Qi et. al. 2003). It did well in capturing the “centroid” of soil-landscape relations, however, it was not able to capture the details of gradations of natural phenomenon.

In this study, the author proposes to represent the soil-landscape knowledge with Gaussian Mixture Models (GMM). The GMM representation has advantage over previous approaches, based on the following three reasons.
Firstly, using a probability distribution to represent the soil-landscape relationships has physical basis. Disregarding the cliffs etc., landscape generally is a continuum. The change of soil properties on the landscape is also in a gradation manner due to the gradual change of terrain variables. Therefore, in the feature space of the environmental variables where a soil is mapped, the soil should appear as a cluster, denser in the middle and sparser towards outside. Figure 1 is a parallel coordinate plot showing the signature profile of several soils across a set of environmental variables. Each color represents one soil type. The cluster center(s) are where all variables favor the soil formation the best and the cluster disperses due to the gradual change of environmental conditions. The plot shows the dimensions independently just for the purposes of easy visualization. The interplay between environmental variables is further discussed in the next paragraph. Therefore, in the perspective of pure data analysis, soil mapping is to find the unique “signature” for each soil series in terms of a set of environmental variables.

**Figure 1 Signature of Different Soils across Environmental Indices**
Secondly, the interactions between environmental conditions are taken into account by GMM representations, rather than assuming each playing a role independently. Joint distribution (multivariate Gaussian distribution) is used to describe the probability of soil formation as a composite effect of the surrounding environmental variables (Eq.1).

\[
p(x) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu)\right)
\]  
\text{Eq. 1}

Where \(X\) is the input environmental vector, \(\mu\) is the vector of mean values, \(\Sigma\) is the covariance matrix of environmental factors, and \(d\) is the total number of dimensions. The interaction between factors is accounted for by the covariance matrix.

Thirdly, mixture model provides a great power of representing a distribution to infinitely approximate the real (Eq. 2). In other words, this nonparametric method enables a close approximation of the real without being restricted to unimodal, or symmetric shaped distributions. This is a great advantage by considering that there might be multiple instances of a soil type on the landscape. That is to say, two or more slightly different sets of environmental conditions might yield the same soils perfectly (as far as human can perceive), or a soil cluster may have multiple “physical centroids” (Figure 2).

\[
p(x | \theta) = \sum_{i=1}^{\epsilon} p(x | \theta(i),i)p(i)
\]  
\text{Eq. 2}
where $P(i)$ is the prior probability of a component and $P(x|\Theta(i))$ is the component density, which is represented as a multivariate Gaussian distribution as in Eq. 1.

Figure 2 Signature of a Soil Type across Environmental Indices

Based on the above three reasons, GMM representations of soil-landscape model are intuitive and suitable. With the GMM representations, Maximum Likelihood (ML) classifier will be used to perform soil mapping. ML evaluates the favorability of the environmental variables for the occurrence of all soil types, and predicates the local soil to be the most likely one.

The rest of the paper is organized as follows. First, the algorithm is discussed briefly. Then a case study is used to illustrate the implementation. Lastly the performance of the algorithm is discussed.
Algorithm

The implementation of the algorithm is based on the ML codes provided on the class webpage\(^1\). The algorithm is described below.

1) Preprocessing data. Preprocessing includes standardization (each feature dimension is scaled independently to a uniform range) and grouping (data on different geology are trained independently);
2) Extracting the features for each class from the training set;
3) Using Expectation Maximization (EM) algorithm to fit a GMM to each class using the extracted samples. The number of components (mixtures) is preset and K-means clustering is used to initialize the cluster centers.
4) Classifying the test data (a new mapping area) with the GMMs based on Maximum Likelihood.

**Figure 3 Pseudo code of Training and Testing Procedures**

<table>
<thead>
<tr>
<th>Standardize feature dimensions of training set</th>
</tr>
</thead>
<tbody>
<tr>
<td>For each geology group in the training data</td>
</tr>
<tr>
<td>For each soil type in the geology group</td>
</tr>
<tr>
<td>Fit a GMM using EM algorithm (# of mixtures is preset, k-means is used to initialize the cluster centers)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Standardize feature dimensions of testing set</th>
</tr>
</thead>
<tbody>
<tr>
<td>For each testing point</td>
</tr>
<tr>
<td>For each class in the corresponding geology group</td>
</tr>
<tr>
<td>Calculate the corresponding likelihood based on GMM</td>
</tr>
<tr>
<td>The point is classified into the class with the maximum likelihood</td>
</tr>
</tbody>
</table>

\(^1\) http://homepages.cae.wisc.edu/~ece539/matlab/index.html
Case study

experiment setup

A case study was performed in southwestern Dane county. The training set is a small watershed named pleasant valley, and the testing set is the surrounding quarter-quad named blue mounds about 3 times in size. Pleasant valley watershed is about 12 km², with moderately undulating terrain as the main features. The training set and the testing set have similar physiographical settings and therefore similar soil series.

The DEM was obtained from the SoLIM lab² of the geography department at UW-Madison. All other terrain indices were calculated with standard algorithms. Geology layer was also obtained from the SoLIM lab. There are 2 major geology types in the study area, with 2 minor ones taking up about 1/10 of the total study area. The study has only focused on the two major geology areas. There are 11 soils in total involved in the study. Among them, 6 soils occur on one type of geology and 3 soils on the other, with 2 more drainage soils that are not constrained by geology types. A SoLIM produced soil map was used as the labels in the training instead of the national soil survey map. The reason is that the SoLIM maps are soil-series based, while the SSURGO³ maps are map units based. All data layers used are attached in appendix I.

data preprocessing

The data have been preprocessed in two ways.

---

² 440 Science Hall, 550 North Park Street, Madison, 53706, WI.
³ Soil Survey Geographic Database, Natural Resources Conservation Services, USDA.
Firstly, because soils can be clearly separated into two geology groups, the data are also partitioned into geology groups for training. Though geology can also be treated as a separate dimension in the training, it is inappropriate to represent a categorical variable with Gaussian curves, which may even confuse the classifier.

Secondly, each feature dimension is scaled independently to a uniform range. Maximum likelihood classifier itself doesn’t require the scaling of feature data. However, the GMM representations do require the standardization of feature dimensions. By standardizing, the variance of each dimension is scaled to about the same order of magnitude. Without standardizing, the inference will be biased toward the feature with small variance due to the use of the covariance matrix in the GMM representation. For example, the range of elevation is between 765-1157 ft, and the range of slope gradient is between 0 – 55%. Without standardization, slope gradient implicitly carries a much bigger significant role in the inference, because the effect of elevation (probability distribution) is spread out over a much larger range. In the case study, all feature dimensions are independently scaled to [-5, 5].

**validation**

The mapping result is evaluated in two steps: the classification accuracy in terms of training data, and mapping accuracies based on an existing soil map and field points.

Confusion matrix is used to assess the training accuracy (against the training set) and testing accuracy (against an existing soil map for blue mounds). Classification accuracy rate is calculated by the ratio of sum of diagonal elements over the total testing set.
Eighty-three field points are available in the blue mounds quarter-quad to access the mapping accuracy.

**Result and Discussions**

The discussion is focused on two aspects: the fitness of knowledge representation and the classification accuracy.

Firstly, the advantage of GMM representations is that it not only explicitly describes the typical cases (centroids), it also well captures the gradations of non-typical cases by taking into consideration of interaction of environmental factors. As a result, it well depicts soils on the landscape continuum. For example, it clearly captures soil Council on foot slope positions, and soil Elbaville on back slope positions (see appendix III for confidence maps for each individual soil), which also comply with expert knowledge very well. As an indication, this tool can be used to mine knowledge from existing soil maps where expert knowledge is not available in other forms, especially when not explicitly documented.

Secondly, the mapping achieves a high accuracy of around 80%. The statistics are given in table 1. Figure 4 gives the plots. It is observed that even though training has a slightly higher accuracy, about 2 - 3% higher in average, the classifier performs decently on the testing data. It proves that the soil-landscape knowledge is well captured by the classifier in general.

It is also observed that the performance of classifier increases with the increase of number of components. Though there are small dips in the two curves for
geology area 2, it doesn’t hurt the overall trend. The increase of number of components increases the power of the model to capture the further detail of the distribution, at an expense of high storage and computational load though. When bigger numbers of mixtures are used, the computation speed is significantly slowed, and with even bigger number the program stops because the computer is out of memory. A solution to reduce computational load is to use diagonal matrix to replace the full covariance matrix, by assuming the independence between dimensions. But further studies should be done to test the performance of the classifier in that case.

<table>
<thead>
<tr>
<th># of mixtures</th>
<th>geology area 1</th>
<th></th>
<th>geology area 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>training</td>
<td>testing</td>
<td>training</td>
<td>testing</td>
</tr>
<tr>
<td>1</td>
<td>70.04</td>
<td>68.07</td>
<td>79.80</td>
<td>77.13</td>
</tr>
<tr>
<td>2</td>
<td>76.66</td>
<td>74.50</td>
<td>78.99</td>
<td>76.84</td>
</tr>
<tr>
<td>4</td>
<td>81.51</td>
<td>79.27</td>
<td>80.03</td>
<td>75.55</td>
</tr>
<tr>
<td>8</td>
<td>83.17</td>
<td>80.12</td>
<td>84.07</td>
<td>79.23</td>
</tr>
</tbody>
</table>

Figure 4 classification accuracy vs. # of mixtures
Based on the 83 field points, 64 points are correctly classified by using the output from the 8 mixture models. The mapping accuracy is 77.1%, which is fairly good compared to the 60% accuracy of traditional soil survey.

In addition, the standardization of data turned out to be very effective. The training accuracy was improved from about 55% to about 80% with standardization of feature dimensions. On the contrary, other preprocessing techniques such as data cleaning required in decision trees won’t be critical to the ML classifier. Although inclusion errors may exist in the training map, the ML classifier is not sensitive to such errors as long as it is not of a huge amount.

Reference


USDA Handbook No. 18, 1993, Soil Survey Manual (for Soil Survey Division Staff), United States Department of Agriculture


Appendix I: data layers used in the case study

training set: pleasant valley

elevation

slope gradient

profile curvature

planform curvature

geology

soil map produced from SoLIM

---

4 The 3D scenes are produced with 3dmapper. www.terrainanalytics.com
testing set: blue mounds quarter-quad

Appendix II: Classification result

Soil map produced from SoLIM

Classification result
(the dark blue areas are not mapped)
Appendix III: confidence maps for individual soils