Unsupervised Learning for Geophysical Data Exploration

Robert Granat[(1)] and Andrea Donnellan[(1)]

(1) Jet Propulsion Laboratory, Pasadena, California (e-mail: granat@aig.jpl.nasa.gov; andrea@aig.jpl.nasa.gov, phone: +1-818-393-5353; +1-818-354-4737, fax: +1-818-393-5244).

Abstract

Unsupervised learning techniques provide a way to investigating scientific data based on automated generation of statistical models that describe the data. Because they do not incorporate a priori information, they can be used as an unbiased method to separate data into distinct types. Thus they can be used as an objective method by which to separate data into previously known classes or to find previously unknown or rare classes and sub-classes of data. Hidden Markov models are one type of unsupervised learning method that are particularly applicable to geophysical systems because they include in the model the time relationship between different classes, or states of the system. We have applied hidden Markov models to scientific analysis of seismicity and GPS data from the Southern California region.Preliminary results indicate that the technique can isolate distinct classes of earthquakes from seismicity data, as well as different modes of ground motion from GPS data.

Introduction

In recent years, computerized analysis techniques have become increasingly popular for use in scientific data analysis (Fayyad, 1996 [1], 1999 [2], Stolorz 1998 [4]). Here we discuss the application of an unsupervised learning technique to data mining of large geophysical data sets. By data mining, we mean the process of extraction of interesting information from the data in cases where: (1) either there is so much data that analysis by hand would be impractical or even intractable, or (2) in cases where trends in the data may be subtle enough to evade the notice of a trained human analyst.

Unsupervised learning, also known as clustering, is one approach to such a task. As the end result of an unsupervised learning algorithm, the data is grouped into several classes such that data belonging one class is similar to other data in that class but dissimilar to data in other classes. So, if each data point is viewed as a vector in feature space, then data points who are members of the same class will appear clustered together in feature space. Clustering is a useful approach to take in data mining tasks because it does not require that any of the data be labeled ahead of time; the algorithm is free to determine which data are related without human supervision. This means that unexpected patterns that might otherwise be overlooked because of bias can be discovered.

Hidden Markov models (HMMs, Rabiner 1989 [3]) are one type of unsupervised learning method that are particularly applicable to geophysical systems because they include in the model the time relationship between different classes, or states of the system. The method assumes that the data was generated by an unknown statistical process which at each point in time can be in any one of a given number of states. Each state is associated with a probability distribution of observable outputs and a set of transition probabilities. These transition probabilities determine the probability that the system will be in each of the possible states at the next point in time, given the current state. An iterative approach

is used to find the system model and state sequence that best explain the observed data. The data can be clustered by using the optimal state sequence: each data vector is labeled according to the system state at the time the data was generated. The advantage of this approach is that the end result says more than which data is related to which other data; it also provides information about the confidence with which each class assignment can be regarded, and about the relationship between classes in time. For this reason, an accurate HMM has the potential to provide valuable predictive information if employed in a real time context.

Algorithm

In order to determine the optimal model and state sequence we employ the standard forward-backward method. However, this method suffers from a local maxima problem; that is, the quality of result is dependent on the initial conditions. To overcome this problem, we observe that the optimization method can be viewed as an variant of the expectation-maximization (EM) algorithm commonly employed in finite mixture modeling. By adapting the deterministic annealing EM method of Ueda and Nakano (Ueda 1998 [5]) to the HMM case, we are able to develop a deterministic annealing HMM (DAHMM) method whose results are largely independent of the initial conditions.

Results

In our preliminary investigations, we applied our DAHMM method to GPS and seismicity data collected in the Southern California region. We present some example preliminary results in the subsequent figures.

Figure 1 shows the results of the DAHMM method applied to GPS data collected in the city of Claremont, California. The data used in the analysis is has three components: east-west displacement, north-south displacement, and vertical displacement. Using a five state model, the method is able to separate the data into distinct classes that correspond to physical events. The states before and after the Hector Mine quake of October 1999 are clearly separated, and distinct in turn from a period in 1998 in which a well caused a change in the vertical direction. Sharp movements in the north-south direction were also isolated as a separate class.

Figures 2-4 show the results of the DAHMM method applied to seismicity data taken from the SCEC catalog. The data used in the analysis had five components: latitude, longitude, depth, magnitude, and time until next event. As well, in this experiment the data was filtered to include only events of magnitude greater than or equal to four.

Figure 2 shows a class of earthquakes which includes several major events, including the Hector Mine and Landers earthquakes.

Figure 3 shows a class of earthquakes of relatively large magnitude and a long time until next event.

Figure 4 shows the transition probabilities for the class portrayed in figure 3.

Note that the next event is most likely to belong to the class of large earthquakes, class 16. While the relationship between these two classes has not yet been fully investigated, these examples do demonstrate the potential of this type of analysis.

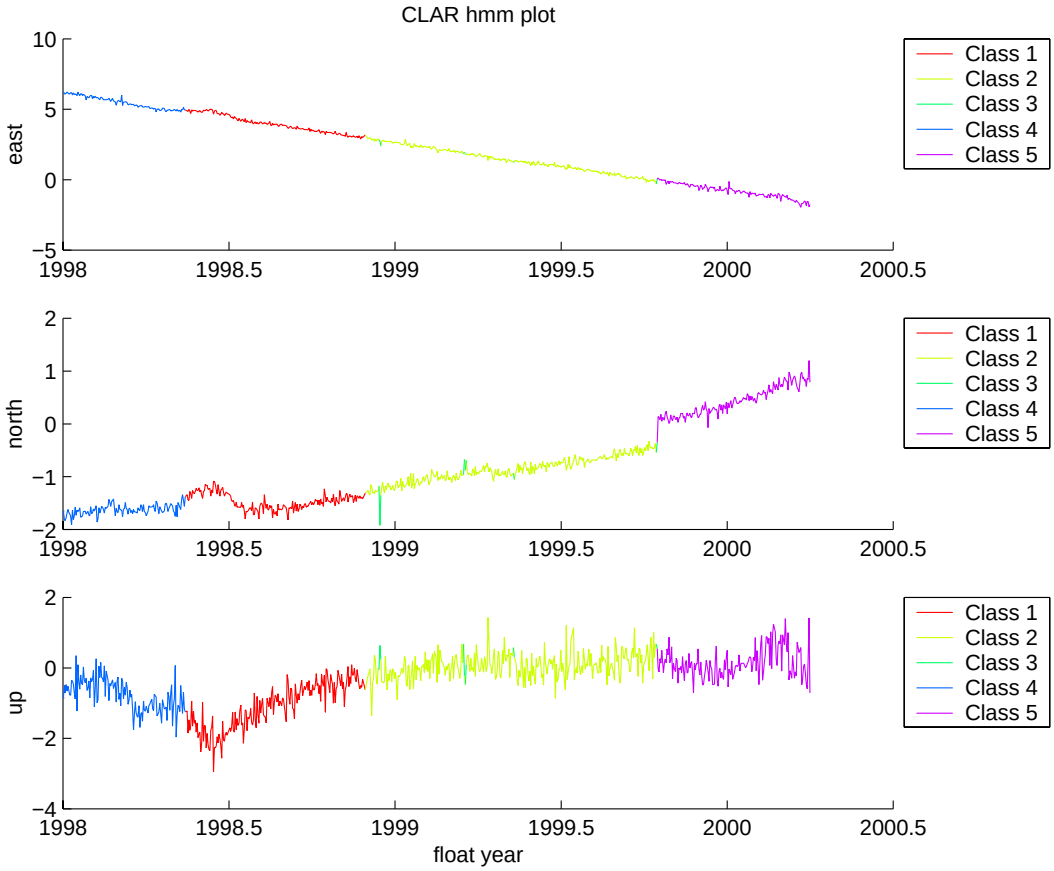

Figure 1: HMM analysis results for Claremont, California GPS receiver.

References

[1] U. M. Fayyad, S. G. Djorgovski, and N. Weir.
From digitized images to online catalogs - data mining a sky survey.
AI Mag., 17(2):51–66, 1996.

[2] U. M. Fayyad and P. Smyth.
Cataloging and mining massive datasets for science data analysis.
J. Comput. Graph. Stat., 8(3):589–610, 1999.

[3] L. R. Rabiner.
A tutorial on hidden markov models and selected applications in speech recognition.
P IEEE, 77(2):257–286, 1989.

[4] P. Stolorz and P. Cheeseman.
Onboard science data analysis: Applying data mining to science-directed autonomy.
IEEE Intell. Syst. App., 13(5):62–68, 1998.

[5] N. Ueda and R. Nakano.
Deterministic annealing em algorithm.
Neural Networks, 11(2):271–282, 1998.

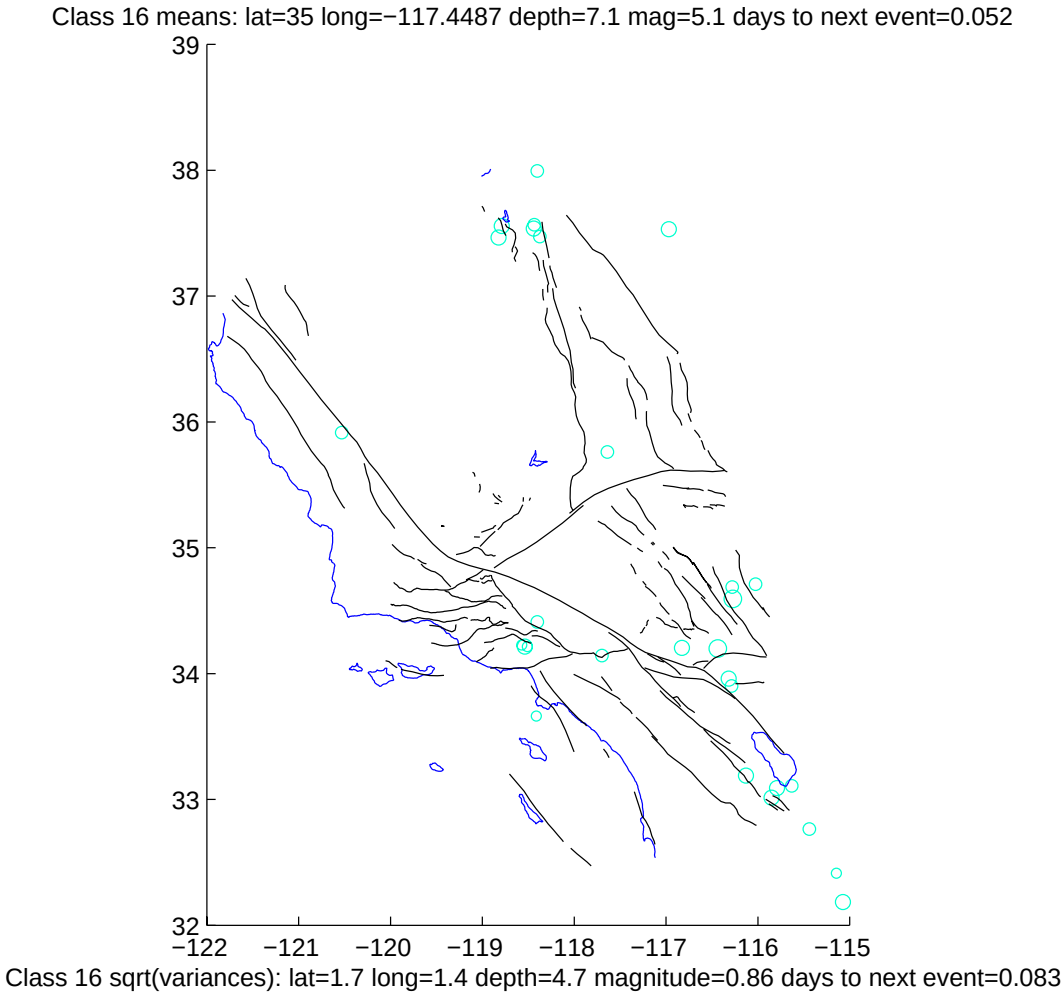

Figure 2: HMM analysis results for Southern California seismicity data.

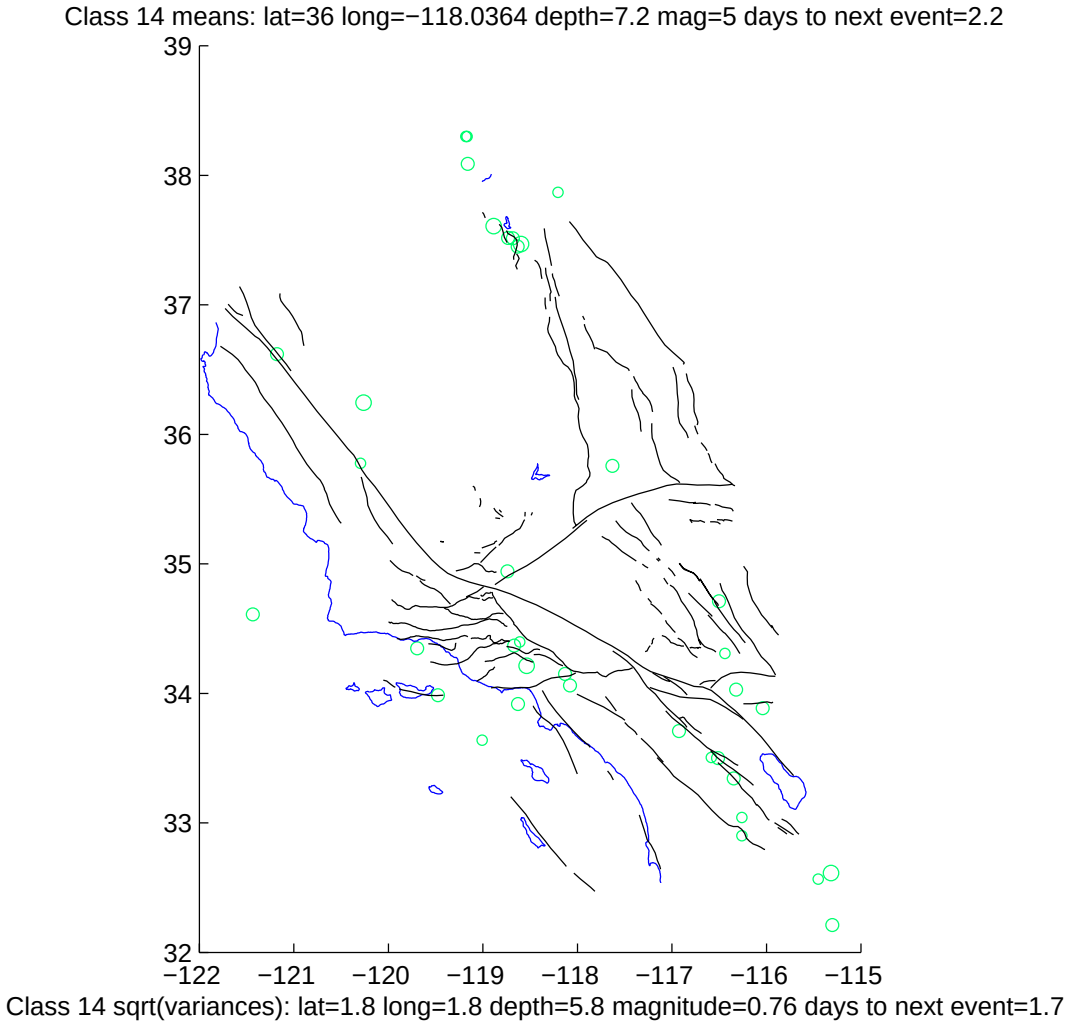

Figure 3: HMM analysis results (transition probabilities for Southern California seismicity data.

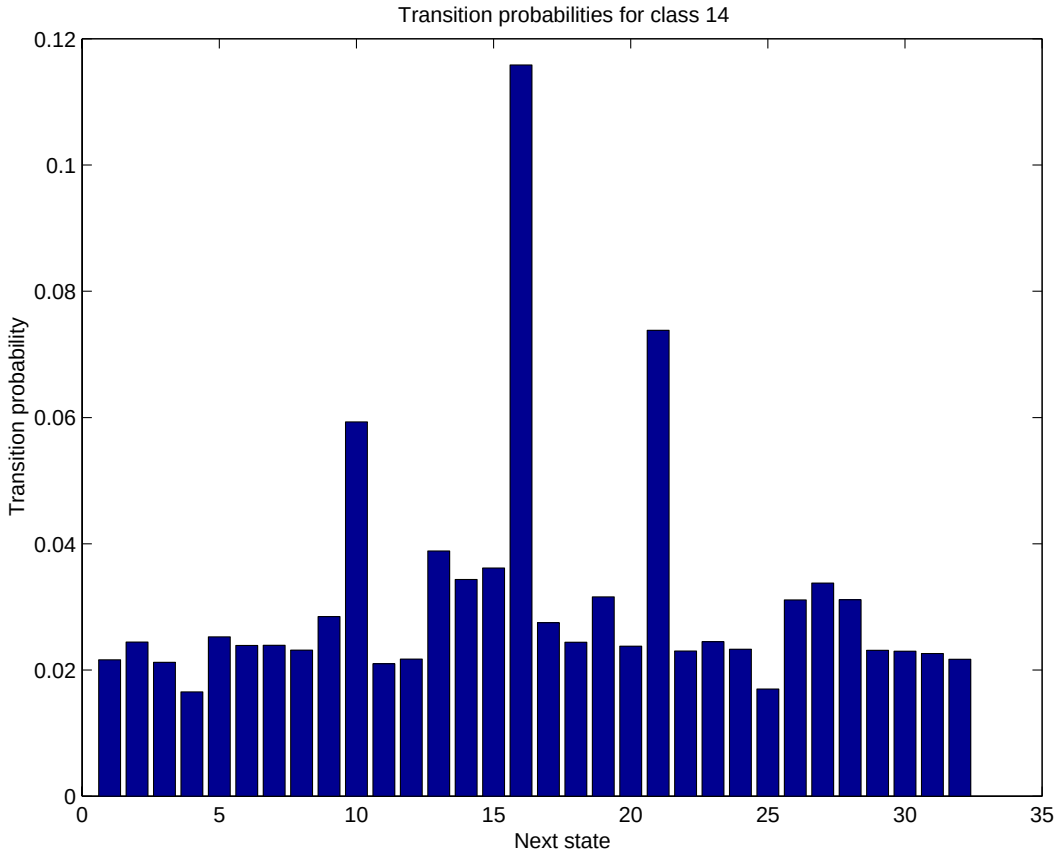

Figure 4: HMM analysis results (transition probabilities for Southern California seismicity data.