Knowledge Discovery From Global Remote Sensing and Climate Data: Results from Supervised and Unsupervised Data Mining

Mark Friedl
Department of Geography and Environment,
Center for Remote Sensing, Boston University
675 Commonwealth Avenue, Boston MA, 02215
friedl@bu.edu

Carla Brodley
Department of Computer Science, Tufts University
Haligan Hall, 161 College Avenue Medford, 02155
brodley@cs.tufts.edu

1. Introduction

This paper describes results and lessons learned from research activities designed to develop data mining and machine learning methods for remote sensing and Earth science data sets. These data sets are acquired by Earth observing instruments onboard polar orbiting satellites, in-situ observations, and model reanalysis and provide a rich source of information related to the properties and dynamics of the Earth's land, oceans, and atmosphere. They are characterized by very large volumes, high dimensionality, and possess both spatial and temporal attributes. The result is a suite of extremely complex, high-dimensional, and heterogeneous data sets that present significant analysis challenges.

To address these challenges we have explored the use of supervised and unsupervised techniques to both extract information and to understand and discover new patterns in these data. In the former case, we have applied ensemble classification methods to the problem of classifying global land cover using data from NASA's Moderate Resolution Imaging Spectroradiometer (MODIS). This is a challenging classification problem that involves large data volumes, a high dimensional and complex set of features, and noisy and often missing data. In the latter case, we have used unsupervised methods including independent component analysis and canonical correlation to identify spatio-temporal patterns of joint ecosystem-climate variability at global scales using gridded climate and remote sensing data sets. Both applications yielded successful results. However, our results also reveal a number of areas that require further research, including both technical issues related to weaknesses inherent to the algorithms, as well as issues related to the application of machine learning algorithms in this domain. In the sections that follow we briefly describe the data sets, applications, and results from our efforts. We conclude with a summary of the main lessons learned from this work and a brief discussion of future areas of research.

2. Data

The data used for this work include four main types:

1. Global 1-km multispectral and multitemporal land surface reflectances from the MODIS instrument onboard NASA's Terra and Aqua spacecraft. Specifically, we have used cloud screened and atmospherically corrected surface reflectances at 16-day intervals in seven wavelength regions in association with 16-day values for the enhanced vegetation index (EVI) for the period from March 2000 to present.
2. A global database of land cover test and validation sites. This database includes 2130 sites (roughly 42,000 km^2) distributed globally that have been characterized in detail with respect to their vegetation, land cover, and surface biophysical conditions via manual interpretation of high resolution imagery performed by analysts.
3. Global data sets of normalized difference vegetation index (NDVI) data from NOAA's Advanced Very High Resolution Radiometer (AVHRR) spanning the period from 1981 to present. These data do not possess the radiometric quality of MODIS, but have the advantage of providing a much longer time series of global observations.
4. Gridded climate data sets related to temperature and precipitation regimes, circulation patterns, and sea surface temperatures, which are a key source of climate variability. These data sets include the 50-year global reanalysis data set, which is produced by the National Centers for Environmental Prediction (NCEP), as well as available reconstructions of global precipitation fields.

3 Analyses

3.1 Supervised Classification of Land Cover from MODIS

For the past eight years we have been developing and testing classification algorithms in support of global land cover mapping for Earth science applications. To do this, we have used a decision tree algorithm (C4.5) in association with boosting. A key input to this activity is the database of land cover test and validation sites described above. We have found this strategy to be quite robust, especially because C4.5 provides effective methods for handling missing data, which are prevalent in MODIS data because of clouds and low-illumination conditions at high latitudes in the winter.

As part of this effort we have developed a number of innovative solutions to improve classifier performance. In particular, we have developed methods to merge information from existing maps (i.e., prior knowledge) via Bayes' rule (McIver and Friedl 2001, 2002). At the same time, a number of challenges remain that we are currently addressing. These issues include: (1) the effect of unbalanced training data on classification results, particularly with respect to minority classes; (2) the need for improved active sampling methods for acquiring new training sites; and (3) feature selection and augmentation.

3.2 Unsupervised Analysis of Joint Variation in Vegetation and Climate

In regards to unsupervised analysis, we have explored a variety of approaches with the goal of uncovering and elucidating relationships between climate patterns and

processes, and ecosystems responses to those patterns. To do this we have tested the use of independent component analysis (ICA), canonical correlation analysis (CCA), and so-called local-linear canonical correlation analysis.

These analyses were performed in three main steps. In the first step, we used ICA to identify unique modes of spatio-temporal variation in time series of NDVI and sea surface temperature measurements from AVHRR. These analyses also revealed subtle artifacts in the data that were previously thought to be removed from the data that were associated with geophysical phenomena (Mount Pinatubo), and instrument changes and calibration (Lotsch et al., 2003). In the second step, we used CCA to examine the joint linear variability between precipitation regimes and ecosystem dynamics at global scales. This analysis revealed regions of join variability that are associated with specific modes of variation in the climate system, particularly the El-Nino Southern Oscillation (Lotsch et al., 2003). In a follow-on study to this work, we again used linear methods, but this time to identify more specific modes of sea surface temperature forcing related to the recent extended northern hemisphere drought (Lotsch et al., 2005).

In the third step of this effort we examined alternative approaches to linear methods such as CCA. Specifically, a key limitation of CCA is that it can only detect linear correlation that is globally valid throughout both data sets. To address this we developed an algorithm that constructs a mixture of local linear CCA models through a process we refer to as correlation clustering (Fern and Brodley, 2003, 2004). In correlation clustering, both data sets are clustered simultaneously exploiting the data's correlation structure such that within any given cluster, the data sets are linearly correlated. Each cluster is then analyzed using traditional CCA to construct local linear correlation models. Results from this algorithm demonstrate that is was able to detect useful correlation patterns, which traditional linear methods such as CCA fail to discover.

4. Discussion and Conclusions

The results from our research over the last several years point to several key conclusions. In regards to supervised classification techniques, a need exists for a more comprehensive approach to both the classification problem and algorithm development. Specifically, most algorithm development, refinement, and implementation activities tend to focus on specific problems (e.g., feature selection), while ignoring the effects of other factors such as the quality and representativeness of the training data. We propose that a key next step will be to develop classification approaches that consider the effects of each stage of the process including training data development and refinement, feature selection, and treatment for infrequent or rare classes that are often penalized by supervised algorithms. Second, in regards to unsupervised methods, a need exists for more computationally efficient algorithms that are able to identify complex, non-linear, and often subtle spatio-temporal relationships among high dimensional remote sensing and climate data sets. The climate and Earth science community is increasingly aware of the complex and non-linear nature of the Earth system. Current tools, however, are not able to perform knowledge discovery from these data sets in a natural and efficient manner. The development of such tools would be a boon to the Earth science community and would almost certainly lead to important new discoveries.

5. References

Fern, X.Z. and C. E. Brodley 2003. Random projection for high dimensional data clustering: A cluster ensemble approach. In Proceedings of the Twentieth International Conference on Machine Learning.

Fern, X.Z and C. E. Brodley 2004. Solving cluster ensemble problems by bipartite graph partitioning. In Proceedings of the Twenty First International Conference on Machine Learning, pages 281–288.

Lotsch, A, M.A. Friedl, and J. Pinzon, 2003. Spatio-Temporal Deconvolution of NDVI Image sequences using independent component analysis, *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 41. No. 12, pp. 2938-2942.

Lotsch, A., Friedl, M.A., Anderson, B.T. and C.J. Tucker 2003. Coupled vegetation-precipitation variability observed from satellite and climate records, *Geophysical Research Letters*, 30(14), 1774, doi: 10.1029/2003GL017506

Lotsch A, Friedl MA, Anderson BT, Tucker CJ, 2005. Response of terrestrial ecosystems to recent Northern Hemispheric drought, *Geophysical Research Letters*, 32 (6): Art. No. L06705.

Mciver, D.K. and M.A. Friedl 2002. Using prior probabilities in decision-tree classification of remotely sensed data, *Remote Sensing of Environment*, Vol. 81, pp. 253-261.

McIver, D.K. and M.A. Friedl 2001. Estimating pixel-scale land cover classification confidence using non-parametric machine learning methods, *IEEE Transactions on Geoscience and Remote Sensing.* Vol 39(9), pp. 1959-1968.