

Machine Learning and Data Mining for Improved Intelligent Data Understanding of High Dimensional Earth Science Data

Prof. Carla Brodley

School of Electrical and Computer Engineering
Purdue University

Prof. Mark Friedl

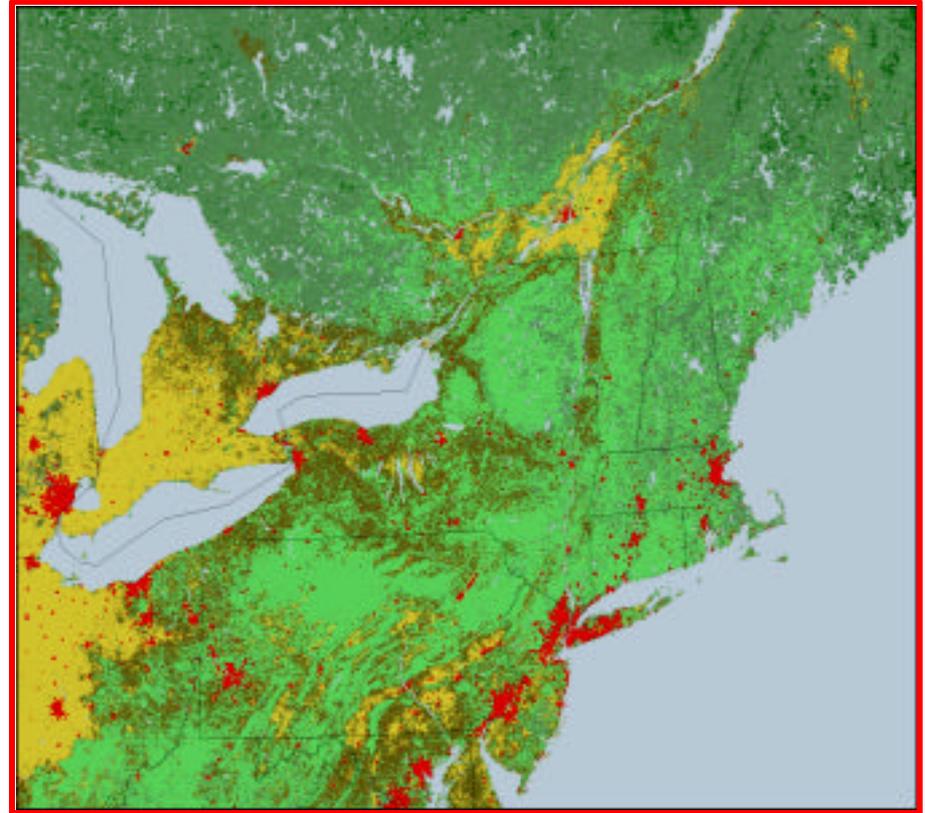
Department of Geography and Center for Remote Sensing
Boston University

Part I: Supervised Learning

- Goal and Technical Objectives:
 - Solve multi-class problems with a large number of classes
 - Develop techniques to improve the classification accuracy when many feature values are missing
 - Improve the minority class performance of the classification system

Technical Problem Statement: Generate Maps of Global Land Cover

- **Large number of classes**
 - 17 classes in MODIS data
- **Cloud cover causes missing feature values**
 - complicates the decision process
 - degrades the classification performance



Sample map for NE US & SE Canada

Technical Problem Statement (Cont.)

- Class distribution of remote sensing data is often highly skewed
 - For example: the smallest class in MODIS data set is 0.8% of the labeled dataset
 - Minority classes are difficult to handle for classifiers
 - A large number of classes is difficult to handle for state of the art data mining methods

Technical Approach

- Recursive Class Elimination (RCE) system for multi-class problems
 - A framework that recursively eliminates unlikely class labels from the label set for a specific test instance
- Lazy decision tree (LazyDT)
 - Builds a customized classifier for each data point to be classifier.
 - Handles missing feature values and improves minority class accuracy
- Lazy Boosting to further improve LazyDT
 - Applies boosting (a state of the art ML technique) to improve the accuracy of LazyDT

Accomplishments and Preliminary Findings

- Accomplishments:
 - Design and implementation of the RCE framework
 - Implementation of LazyDT algorithm
 - Design and implementation of Lazy Boosting
- Preliminary findings:
 - RCE system can improve minority class accuracy with no major sacrifice in overall accuracy (Zhang et al. 2002)
 - LazyDT often outperforms C4.5 on minority classes and when missing feature values abound
 - Lazy Boosting further reduces the error rate of LazyDT (Zhang et al. 2002)

Part II: Unsupervised Learning

- Goal and Technical Objectives
 - Develop methods and techniques to aid information extraction from high data volume remote sensing missions in support of NASA's Earth Science Enterprise
 - Focus on analysis of space-time patterns in high dimensional large volume geophysical data sets
 - Focus on grand challenges in Earth System Science
 - Climate teleconnections, biosphere-atmosphere coupling, ocean-climate feedbacks

Technical Problem Statement

- NASA is currently archiving ~250GB/year in Earth Science data
 - Remote sensing
 - Model output
 - Largely unexplored
- New techniques and methods required to identify and extract interesting and/or anomalous space-time patterns.

Technical Approach

- Application of unsupervised methods to ocean, atmosphere and remote sensing data sets
 - Sea surface temperature time series, NDVI time series, atmospheric time series
- **Independent Component Analysis**
 - New technique, similar in principal to empirical orthogonal (EOF) function analysis
 - More flexible, reveals *uncorrelated and independent* components
 - Yields spatial and temporal patterns

Data and NASA Relevance

- Data sets derived from or relevant to NASA's Earth Science Enterprise
 - Satellite remote sensing (Terra)
 - Climate model output
 - Long term atmospheric observations
- Questions specifically designed to address NASA's need for efficient methods of identifying and extracting significant patterns in large volume, high dimensional Earth science data sets.

Accomplishments and Preliminary Findings

- Accomplishments:
 - Compilation of test data sets
 - Implementation and testing of ICA algorithm
- Preliminary findings:
 - Isolation of previously undetected and distinct La Nina and El-Nino signatures in SST ICA (Lotsch et al. 2002)
 - Isolation of unique spatio-temporal dynamics in large scale vegetation response to decadal scale climate forcing in Africa based on long time series of NDVI data (Lotsch and Friedl, 2002)

Technical Significance and Expected Impact on NASA (Supervised and Unsupervised Research)

- Adaptation of innovative new data mining methods to space-time data sets for climate and geophysical data.
 - Results from this effort will provide Earth Science Enterprise activities with a new method for examining and understanding large volume, high dimensional remote sensing and geophysical datasets
 - Development of new data mining methods for handling multiple classes, minority classes and missing data
 - Results from this effort will provide better methods for generating accurate, complete, global land cover maps.

Personnel

- Carla Brodley, Co-PI
- Mark Friedl, Co-PI
- Alex Lotsch, Ph.D. student, Boston Univ.
- Su-Yin Tan M.A. student, Boston Univ.
- Xiaoli Zhang, Ph.D. student, Purdue Univ.

References

- Lotsch, A., M.A. Friedl and B.E Anderson 2002. Identification of non-linear sea surface temperature patterns using temporal independent component analysis, in preparation for *Journal of Climate*, expected submission, June 2002.
- Lotsch, A. and M.A. Friedl 2002. Unmixing NDVI image sequences using spatial and temporal independent component analysis, in preparation for *IEEE Transactions on Geoscience and Remote Sensing*. expected submission, August 2002.
- Zhang, X. and Brodley, C. Local Boosting Algorithms, in preparation of *Journal of Machine Learning Research*. expected submission, June 2002.
- Zhang, X., Brodley, C. and Friedl, M. Generating Global Land Cover Maps: Handling cloud cover and underrepresented classes, in preparation for *IEEE transaction on Remote Sensing*. Expected submission, October 2002.