**High-performance Spatiotemporal Data Mining**
Ranga Raju Vatsavai, Varun Chandola
GIST/CSED, Oak Ridge National Laboratory,
Oak Ridge, TN 37831, USA

**Goal**
- Identify and characterize changes
- Phenological (crop growth)
- Land-use (crop type)
- Develop analytical capabilities to monitor crop biomass at regional and global scales
- Develop scalable solutions

**Challenges**
- Large spatial and temporal extents
- High volume of data throughput
- Spatial homogeneity
- Spatial heterogeneity
- Multimodal data
  - Multi-resolution (spatial and temporal)
  - Multi-sensor
  - Overlapping data products
  - Noise (such as cloud cover)
- Natural changes vs. real changes
  - Crop phenology (changes in NDVI are expected due to growing nature of crops)
  - Temporary changes (rain, snow, dry)
- Insufficient Ground-truth data
- Aggregate vs. sub-classes

**Semi-supervised Learning**
- Semi-supervised Learning
- Gaussian Mixture Model (GMM)

\[ p(x_i | \theta) = \sum_{j=1}^{M} \alpha_j p_j(x_i | \theta_j) \]

- Estimate GMM parameters using EM by posing unlabeled sample as missing data
- EM essentially consists of following two steps:
  - E-Step:
    \[ \epsilon_j = \frac{1}{\sum_{i=1}^{N} \epsilon_i} \exp \left( -\frac{1}{2} (x_i - \mu_j)^T \Sigma_j^{-1} (x_i - \mu_j) \right) \]
  - M-Step
    \[ \alpha_j = \frac{\sum_{i=1}^{N} \epsilon_i}{N} \]
    \[ \mu_j = \frac{\sum_{i=1}^{N} \epsilon_i x_i}{\sum_{i=1}^{N} \epsilon_i} \]
    \[ \Sigma_j = \frac{\sum_{i=1}^{N} \epsilon_i (x_i - \mu_j) (x_i - \mu_j)^T}{\sum_{i=1}^{N} \epsilon_i} \]

**Gaussian Process Change Detection**
- Generalization of multivariate Gaussian distributions
- Captures non-linear dependencies
- Non-parametric technique
- Computationally expensive
- Extended for change detection
- Designed a new covariance function

**Gaussian Process Classification**
- Signature Extension is a problem when dealing with large geographic region classification
- In practice, it is generally not possible to have data from all possible locations
- Train separate models for each location – needs lot of data
- Train one model, use on all – poor performance
- Train on some data, adapt for others – needs adaptive model

**Multisource Data Fusion**
- SSL for Multisource Data
  - Mixture of Mixtures Distributions

\[ p(x_i | \theta) = \sum_{j=1}^{M} \alpha_j \prod_{k=1}^{K} p_j(x_{ik} | \theta_k) \]

- Two models (l=2)
  - Continuous Distribution (for image data)
  - Discrete distribution (for categorical variables – ancillary geospatial data)

**References**

**Performance Results**
- Change of distribution over space is modeled by
  \[ p(x_i | y) \sim N(\mu, \Sigma) \]
  \[ p(x(s) | y) \sim N(\mu(s), \Sigma(s)) \]

**Spatially Varying Data**

**Acknowledgements:** Prepared by Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, Tennessee 37831-6285, managed by UTBattelle, LLC for the U.S. Department of Energy under contract no. DEAC05-00OR22725. Research was supported through LDRO program
Contact: Ranga Raju Vatsavai (vatsavairr@ornl.gov)