Local Learning on High Dimension, Imbalanced, and Noisy Data:

A Framework for Long-Lead Extreme Precipitation Clusters Forecasting

Dawei Wang[1], Wei Ding[1], Yang Mu[1], David L. Small[2], Shafiqul Islam[2]

[1] Department of Computer Science, University of Massachussets Boston

[2] Department of Civil and Environmental Engineering, Tufts University

Abstract:

Extreme Flood is usually a consequence of a sequence of precipitation events occurring over from several days to several weeks. Certain atmospheric regimes (e.g., blocking) can lead to sequence of precipitation events. However, the task of long-term (5-15 days) forecasting of precipitation clusters will suffer from overwhelming number of relevant features and high imbalanced, multimodal, and noisy sample sets. Existing atmospheric models which rely on nonlinear deterministic systems cannot deal with such a huge feature space to provide accurate long-range predictability of weather. In this work, we develop a data mining framework for long-lead extreme precipitation cluster forecasting through the identification of atmospheric regime precursors. We synthesize a representative feature set that describes the atmosphere motion, and then we design a novel *Bi-Class Streaming Feature Selection (BCSFS)* algorithm for feature selection on imbalanced data. After the dimension reduction step, we apply the *Local Discriminative Distance Metrics Ensemble Learning (LDDM)* algorithm, which learns distance metrics according to different training samples and predicts a test sample by classifiers ensemble, to learn from the multimodal and noisy sample set and make prediction through classification. An extensive empirical study is conducted on historical precipitation and associated flood data collected in the State of Iowa.