

Supervised Learning from Very Large, High Dimensional Remote Sensing Data Sets

Mark A. Friedl (Department of Geography and Center for Remote Sensing, Boston University)

`friedl@bu.edu`

Carla Brodley

Abstract

In recent years machine learning and data mining methods have become increasingly common in remote sensing applications. One area in which such techniques are particularly useful is classification of remotely sensed data for land cover and vegetation mapping applications. In this paper we describe the techniques and algorithms being used to map global land cover using data from the Moderate Resolution Imaging Spectroradiometer (MODIS) onboard NASA's Terra spacecraft. Data provided by MODIS include global multispectral observations in seven wavelength bands acquired at 16-day intervals over 12 month periods. The spatial resolution of these data is 1 km. Thus, this classification problem involves very large data volumes with high dimensionality (roughly 100 GB and 180 features). The classification algorithm uses a supervised approach. Training data are provided by a database of over 1000 representative land cover sites that have been compiled from high-resolution satellite data globally. Because of the diversity of global land cover and the complexity of the feature space provided by MODIS, common classification methods do not work well for this problem. To provide robust, repeatable, and accurate maps of land cover at global scales a variety of data mining and machine learning approaches have been utilized. Specifically, we describe techniques to filter training data, include contextual domain information derived from existing land cover maps, and novel uses of ensemble classification methods for this problem domain. Sample results from these algorithms will be presented based on recently available data from MODIS.