Project Proposal

Hyperspectral Data Classification

Team members names
Purdue University
wang1766@purdue.edu, yyy@purdue.edu, zzz@purdue.edu

October 1, 2015

1 Introduction

Conservation tillage management has been advocated for the purpose of soil preservation and sustainable crop production. Conservational tillage practice induces less surface disturbance and leaves more crop residues, which can decrease runoff rate, improve soil and water quality, and increase organic matter. The demand for the mapping of crop tillage practices has been brought up for precision agricultural management and appraisal. However, current methods for mapping massive crop tillage practices are mainly done with field investigations, which are labor costing, time consuming, subjective, and make it difficult to generate widely distributed survey data. Remote sensing technology provides a more rapid, accurate, and objective solution. Moreover, the vast data from remote sensing in agriculture require more efficient approaches in data analytics, including tillage mapping, in support of management decisions.

Recently, hyperspectral remote sensing has gained attention in the remote sensing application community. Hyperspectral imaging generates hundreds of images corresponding to different wavelength channels for the same area on the surface of the earth. A hyperspectral image is a 3-D cube of data with the width and length of the array corresponding to the spatial dimensions and the continuous spectrum of each point as the third dimension, which enables discrimination of materials based on their spectral characteristics. One of the most important applications of hyperspectral data is image classification, where pixels are labeled to one of the classes based on their spectral characteristics.

However, due to the large amount of data, high correlation between bands, directly conducting the classification not only results in slow classification speed but also low classification accuracy. Besides, hyperspectral data also present difficult challenges for supervised statistical classification, where labeled training data are used to estimate the parameters of the conditional probability density functions. In fact, the dimensionality of the data is high while the quantity of training data is often small. Taking into account of these factors, the feature extraction is often conducted to reduce the dimension of hyperspectral image prior to the classification.

Motivation

2 Related work

Describe any previous work you found related to your project

3 Problem Formulation

The hyperspectral image classification task is perfectly suitable to be modeled as a machine learning problem. Each pixel of the hyperspectral image is an example in this problem. For a typical 145*145 high dimension image, we will have 21025 examples in total, which will be split into training datasets, validating datasets and testing datasets later.

The objective: Build a multi-class classifier, where the image data associate with each pixel is the input and there are 16 labels corresponding to 16 classes as the output.

Inputs objects: The image data associate with each pixel is the input. For each example, there are 200 attributes with each attribute varies from 0-1. Labels: There are 16 classes of the pixel, so we have 16 labels with each corresponding to one class.

Possible features: Our dataset is unbalanced, the size of labeled samples is very large or small for certain few classes. So principal component analysis may be needed before building the learning model. Moreover, since our data's dimension is very high, dimension reduction process may applied in the beginning.

4 Data and Evaluation Plan

The data set we plan to use is from a hyperspectral image that was taken by a drone. The whole data set contains all the pixels within the image, which will be further divided into training set and testing set. The whole data set contains 21025 rows, 200 columns of features and 1 column of label. All the feature values were normalized and ranged from 0 to 1000. The label column contains 16 different values. In other words, we have a multi-classification problem.

Below is a pilot evaluation plan:

- Data Check
  The validity of the whole data set will be checked. Any missing values, non-numeric values and outliers will be identified.

- Data split
  The whole data set will be split into training and testing set. Because the data set is unbalanced (the smallest label group contains 20 examples, while the largest label group contains more than 2000 examples), up-sampling of small label groups or down-sampling of the large label groups will be conducted when the training data is selected. In addition, the training set will be further randomly divided into 5 folds of equal size, in order to facilitate the following cross validation process.

- Feature Selection
  Considering the dimensionality of the data set is relatively high, Principal component analysis

(PCA) will be used to conduct dimensionality reduction. We plan to keep at least 90% of the original variance, and the principal components will be used in the following data analysis.

- Model Selection
  Considering our task is classification based on numeric features, we plan to construct models using the following algorithms:
  - Random Forest
    Hyper-parameter to be tuned: Randomly selected features, maxiumn number of trees. This model is available in the Sklearn package of Python.
  - Neural Network
    For this model, BackpropTrainer algorithm from Pybrain package of Python will be used. The hyper-parameters include: the hidden layers, learning rate, batch size and so on.
  - Linear SVM For this model, the tunable hyper-parameters include cost and maximum iterations. This model is available in the Sklearn package of Python.
  - Kernel SVM If necessary, different kernel functions will test. The available kernels in the Sklearn package include: linear, polynominal, rbf, sigmoid and precomputed.
- Model evaluation
  After the construction of the above models, the performance of the different models will be evaluated by using 5-fold cross validation. Since the correct label of all data points are known, misclassification rate will be used as the subjective function.
- Model testing
  For the models with satisfied performance, the error rate will be tested against the held out testing data.