COMPARATIVE ANALYSIS OF DATA MINING APPROACHES IN REMOTE SENSING

Ranga Raju Vatsavai, Budhendra Bhaduri*

Oak Ridge National Laboratory, PO.Box 2008, TN, USA

1. INTRODUCTION

In this article we describe five major classification approaches and provide a comparative analysis on a multispectral image. First we define the classification problem as following. Given a sample set of input-output pairs, the objective of supervised classification is to find a function that learns from the given input-output pairs, and predicts an output for any unseen input (but assumed to be generated from the same distribution), such that the predicted output is is as close as possible to the desired output. The name "supervised" comes from the fact that the input-output example pairs are given by an expert (teacher). We now describe five major classification schemes.

2. MAJOR CLASSIFICATION SCHEMES

2.1. Maximum Likelihood Classifier

Maximum likelihood classification is one of the most widely used parametric and supervised classification technique in remote sensing field [1, 2]. Assuming that sufficient ground truth (training) data is available for each thematic class, we can estimate the probability distribution $p(x|y_i)$ for a class (y_i) that describes the chance of finding a pixel from that class at the position $\mathbf{x}$. This estimated $p(y_i|x)$ can be related with the desired $p(x|y_i)$ using Bayes' theorem: $p(y_i|x) = \frac{p(x|y_i)p(y_i)}{p(x)}$, where $p(y_i)$ is the probability that class y_i occurs in the image, also know as 'a priori' probability, and $p(\mathbf{x})$ is the probability of finding a pixel from any class at location $\mathbf{x}$. By assuming multivariate normal model for class probability distributions, the discriminant function $g_i(x)$ for the maximum likelihood classification can be written as the following: $g_i(\mathbf{x}) = \ln p(\omega_i) - \frac{1}{2}(\mathbf{x} - \mathbf{m}_i)^t \Sigma_i^{-1}(\mathbf{x} - \mathbf{m}_i)$.

2.2. Logistic Regression

Given an $n-$vector $\mathbf{y}$ of observations and an $n \times m$ matrix $\underline{X}$ of explanatory data, classical linear regression models the relationship between y and $\underline{X}$ as $\mathbf{y} = \mathbf{X}\beta + \epsilon$. Here $\mathbf{X} = [1, \underline{X}]$ and $\beta = (\beta_0, \ldots, \beta_m)^t$. The standard assumption on the error vector ϵ is that each component is generated from an independent, identical, zero-mean and normal distribution, i.e., $\epsilon_i = N(0, \sigma^2)$. When the dependent variable is binary, the model is transformed via the logistic function and the dependent variable is interpreted as the probability of finding a given class. Thus, $Pr(l|y) = \frac{e^y}{1+e^y}$. This transformed model is referred to as **logistic** regression [3]. Binary classification can be extended to multi-class classification, using one-against-others approach.

2.3. Decision Trees

A decision tree (DT) is a supervised classifier that recursively partitions a data set into smaller subdivisions based on a set of simple tests at each internal node in the tree. The leaf nodes represents the class labels y_i. Training data set is used to learn the split conditions at each internal node and to construct a decision tree. For each new sample (i.e., feature vector x), the classification algorithm will search for the region along a path of nodes of the tree to which the feature vector x will be assigned. That is, the classification of a region is determined by a path from the root node to a leaf node. Performance of three different types of decision trees, namely, univariate, multivariate, and hybrid decision trees for remote sensing data classification, against MLC were reported in [4]. In this study, we used C4.5 [5], a publicly available univariate decision tree software.

*Contact Author (vatsavairr@ornl.gov)

2.4. Neural Networks

Artificial neural networks, which are non-parametric classifiers as opposed to Bayesian classifiers, are gaining popularity in remote sensing image classification. This popularity can attributed to several factors: 1) previous studies [6] have shown that their performance is as good as MLC and in many cases even better accuracy, 2) they are non-parametric, so they are capable of classifying multi-source data, where as parametric classifiers have problems with multi-source data, and 3) they have several desirable characteristics like nonlinearity, adaptability and fault tolerance. The back-propagation algorithm is the most common method of training multi-layer feed-forward neural networks (also know as multi-layer perceptrons).

2.5. Support Vector Machines

Recently, support vector machines (SVMs) have proven to be very useful in remote sensing image classification []. Though SVMs were originally developed for binary classification problems, it can be easily extended to multi-class classification using one-against-others approach.

3. RESULTS

We applied all the five classification schemes on Cloquet site encompasses Carlton County, Minnesota, which is approximately 20 miles southwest of Duluth, Minnesota. The region is predominantly forested, composed mostly of upland hardwoods and lowland conifers. We used a spring Landsat 7 scene, taken May 31, 2000, and clipped to the study region. The final rectified and clipped image size is 1343 lines x 2019 columns x 6 bands. The training data set consisted of 60 plots and independent test data set consisted of 205 plots. Table 1 summarizes the individual class and overall classification accuracies of all five classification schemes described in this paper.

	X	MLC	Log.Reg	C4.5	Neural.Net	SVM
1	Hardwood.1	81.15	71.01	85.00	88.40	86.11
2	Hardwood.2	90.63	86.71	76.50	91.70	90.11
3	Conifer	79.31	90.41	90.40	85.80	99.01
4	Agriculture	94.01	97.42	88.90	82.90	63.91
5	Urban	99.99	99.99	95.60	100.00	81.23
6	Wetlands	87.52	77.91	67.70	81.90	99.21
7	Water	66.66	99.99	70.40	100.00	98.24
8	Overall	86.94	84.77	78.48	86.72	84.81

Table 1. Comparision of Classification Accuracies

4. REFERENCES

[1] M. Hixson, D. Scholz, and N. Funs, "Evaluation of several schemes for classification of remotely sensed data," *Photogrammetric Engineering & Remote Sensing*, vol. 46, pp. 1547–1553, 1980.

[2] A.H. Strahler, "The use of prior probabilities in maximum likelihood classification of remote sensing data," *Remote Sensing of Environment*, vol. 10, pp. 135–163, 1980.

[3] L Anselin, *Spatial Econometrics: methods and models*, Kluwer, Dordrecht, Netherlands, 1988.

[4] Mark A. Friedl and Carla E. Brodley, "Decision Tree Classification of Land Cover from Remotely Sensed Data," *Remote Sensing of Environment*, vol. 61, pp. 399–409, 1997.

[5] J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Diego, CA, 1993.

[6] A. Benediktsson, P.H. Swain, and O.K. Ersoy, "Neural network approaches versus statistical methods in classificaion of multisource remote sensing data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 28, no. 4, pp. 550, 1990.