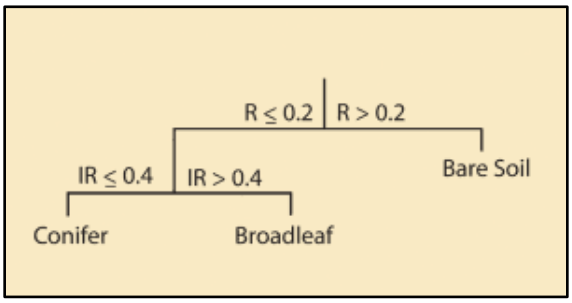
Classification Tree Analysis

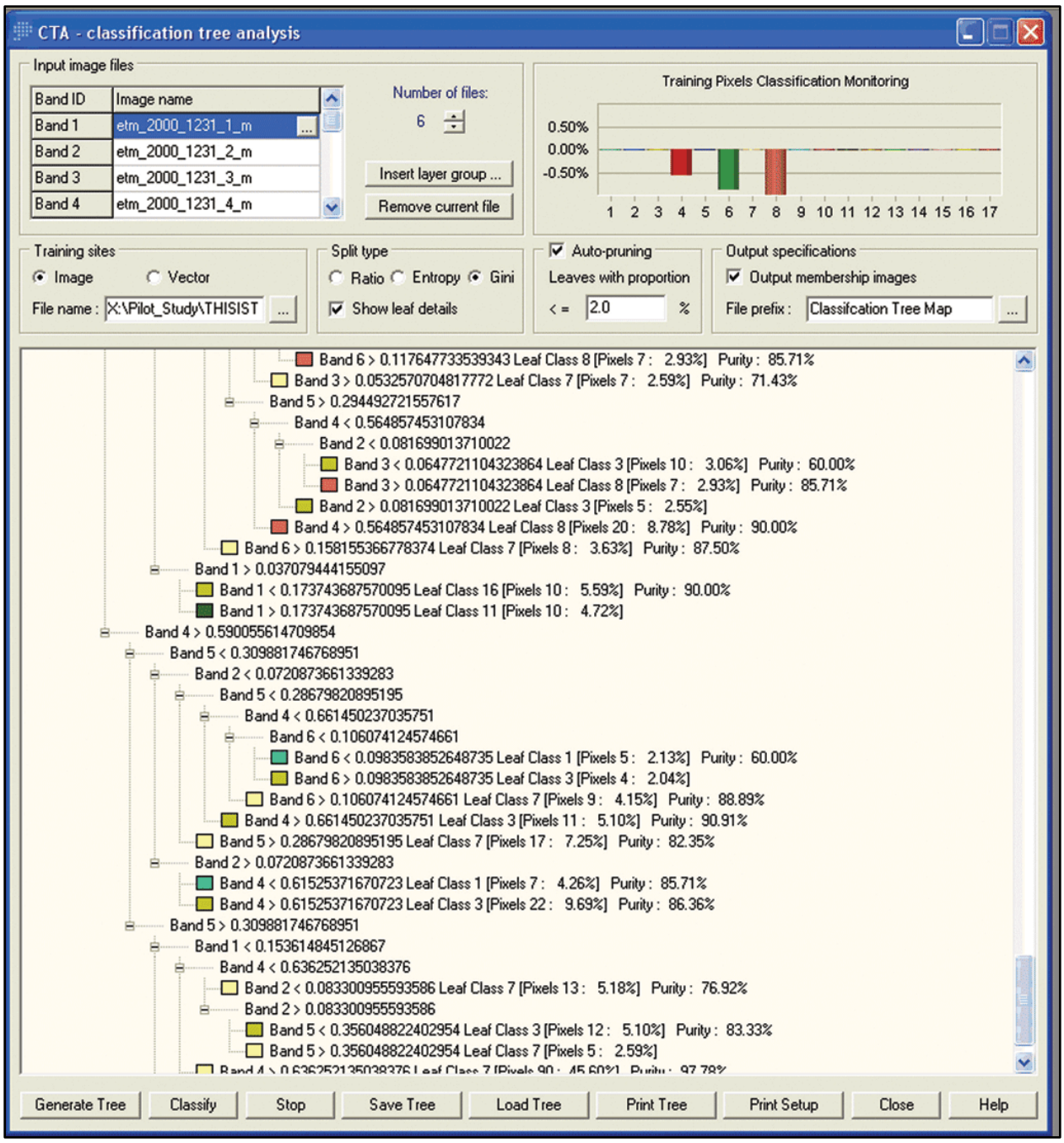
By Clark Labs

Classification Tree Analysis (CTA) is a type of machine learning algorithm used for classifying remotely sensed and ancillary data in support of land cover mapping and analysis. A classification tree is a structural mapping of binary decisions that lead to a decision about the class (interpretation) of an object (such as a pixel). Although sometimes referred to as a decision tree, it is more properly a type of decision tree that leads to categorical decisions. A regression tree, another form of decision tree, leads to quantitative decisions.



A classification tree is composed of branches that represent attributes, while the leaves represent decisions. In use, the decision process starts at the trunk and follows the branches until a leaf is reached. The figure above illustrates a simple decision tree based on a consideration of the red and infrared reflectance of a pixel.

Classification Tree Analysis (CTA) is an analytical procedure that takes examples of known classes (i.e., training data) and constructs a decision tree based on measured attributes such as reflectance. In TerrSet the CTA module is based on the C4.5 algorithm. In essence, the algorithm iteratively selects the attribute (such as reflectance band) and value that can split a set of samples into two groups, minimizing the variability within each subgroup while maximizing the contrast between the groups.



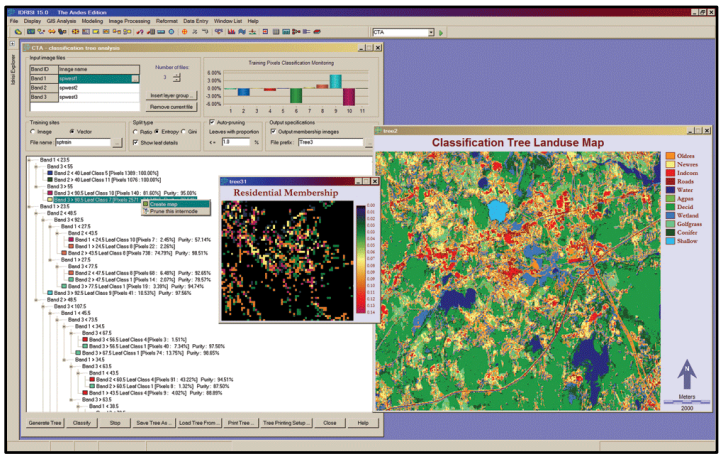
A set of tools is included for training and pruning a classification tree. The monitoring graph shows the incremental inclusion of the training pixels per class during the analysis. Once the analysis is complete, the proportion of misclassified training site pixels will be displayed as well as a tree showing the splits for the relevant input images and the split threshold. Leaf details include the number of pixels in a leaf, the percentage of these pixels that belong to the total pixels in that class, and a purity index that is the percentage of correct pixels belonging to the specified class in that leaf to the total number of pixels in the leaf.

Because it can take a set of training data and construct a decision tree, Classification Tree Analysis is a form of machine learning, like a neural network. However, unlike a neural network such as the Multi-Layer Perceptron (MLP) in TerrSet, CTA produces a white box solution rather than a black box because the nature of the learned decision process is explicitly output. This is one of the main attractions of CTA. The structure of the tree gives us information about the decision process.

Other attractions are that it is simple to understand and it is non-parametric-it does not require that the data associated with a particular class on a particular attribute follow any specific distribution (such as a normal distribution). Thus, for example, it is capable of handling a class with unusual characteristics such as impervious surfaces, which contain both low (asphalt) and high (concrete) reflectors.

As with all classifiers, there are some caveats to consider with CTA. The binary rule base of CTA establishes a classification logic essentially identical to a parallelepiped classifier. Thus the presence of correlation between the independent variables (which is the norm in remote sensing) leads to very complex trees. This can be avoided by a prior transformation by principal components (PCA in TerrSet) or, even better, canonical components (CCA in TerrSet). However, the tree, while simpler, is now more difficult to interpret.

The second caveat is that, like neural networks, CTA is perfectly capable of learning even non-diagnostic characteristics of a class as well. For example, if we were using CTA to learn how to distinguish between broadleaf and conifer forest, and if our training sample for broadleaf included some gaps with an understory of grass, then all grass areas would be classified as broadleaf. Thus CTA includes procedures for pruning meaningless leaves. A properly pruned tree will restore generality to the classification process.



The CTA module provides hard and/or soft classified output maps. There is one soft output for each class. Each pixel in a soft output image is associated with a degree of membership for the

class at that THE TRAINING particular leaf it was classified from. If a pixel is not associated with that class, it will be assigned a zero.

The Training

The user must first use the training samples to grow a classification tree. This is called the training step. Then, the whole image is classified using this tree.

To start, all of the training pixels from all of the classes are assigned to the root. Since the root contains all training pixels from all classes, an iterative process is begun to grow the tree and separate the classes from one another. In Terrset, CTA employs a binary tree structure, meaning that the root, as well as all subsequent branches, can only grow out two new internodes at most before it must split again or turn into a leaf. The binary splitting rule is identified as a threshold in one of the multiple input images that isolates the largest homogenous subset of training pixels from the remainder of the training data.

The tree grows by recursively splitting data at each internode into new internodes containing progressively more homogeneous sets of training pixels. A newly grown internode may become a leaf when it contains training pixels from only one class, or pixels from one class dominate the population of pixels in that internode, and the dominance is at an acceptable level specified by the user. When there are no more internodes to split, the final classification tree rules are formed.

The Classification

The second step of the CTA technique is image classification. In this step, every pixel is labeled with a class utilizing the decision rules of the previously trained classification tree. A pixel is first fed into the root of a tree, the value in the pixel is checked against what is already in the tree, and the pixel is sent to an internode, based on where it falls in relation to the splitting point. The process continues until the pixel reaches a leaf and is then labeled with a class.

Facilitated by an intuitive graphical display in the interface, the classification rules from the root to a leaf are simple to understand and interpret. Input images can be numerical images, such as reflectance values of remotely sensed data, categorical images, such as a land use layer, or a combination of both.

If it is known that a data set obeys a certain distribution pattern, you may want to use a proper parametric classifier other than the classification tree approach. For example, if it is known that the image data obey the Gaussian distribution, a parametric classifier, such as MAXLIKE in TerrSet, may be preferred.