Chapter 6

Classification Tree

TREE-LIKE CLASSIFIER

Classification tree is a part of tree-like classification algorithms broadly used in machine learning, data mining, statistics and their applications. The book (Breiman, Friedman, Olshen, & Stone, 1984), covering not only classification but also regression trees, describes this type of classifier in detail. In fact, the full name – CART – which stands for ‘Classification and Regression Trees’, is frequently used to denote this algorithm. Brief history, preceding the emergence of CART, is presented in (Steinberg, 2009). Due to its popularity and simple interpretation of how classification was done, CART is described in many textbooks (Martinez & Martinez, 2002), (Hastie, Tibshirani, & Friedman, 2003), (Russell & Norvig, 2003), (Kuncheva, 2004), (Bishop, 2006), (Ripley, 2007), (Rokach & Maimon, 2008), (Izenman, 2008). The related algorithms to CART are ID3, C4.5 and C5.0 (Quinlan, 1993), which are not discussed in this book but their description can be found, for example, in (Quinlan, 1993), (Wu & Kumar, 2009).

The CART decision tree is a binary multi-stage decision procedure. The tree is made up of nodes and branches. Nodes can be either internal or terminal. The
terminal nodes are also called leaves. Internal nodes split into two children. These children can be 1) both internal nodes, 2) both terminal nodes or 3) one child can be terminal while the other can be internal. Terminal nodes (leaves) do not have children. A terminal node has a class label associated with it, such that all instances falling into this node are assigned to that class. A single-split tree with only two terminal nodes is called a stump. That is, the stump divides the data into two classes based on a single feature value.

Why binary splits? Rather than splitting each node into two nodes, we could consider multiway splits into more than two groups. However, multiway splits are not very good strategy as they fragment the data too quickly, leaving insufficient data for the next level. Moreover, any multiway split can be represented as a sequence of binary splits so there is no reason to prefer multiway splits over binary ones.

In tree construction the data are handled in their raw form as feature vectors. In the root node the entire data set is split into two children nodes. Each of the children is then further split into grandchildren nodes that are in turn split to obtain great-grandchildren, etc., i.e. the original data set is recursively split between tree nodes as the tree grows.

The tree is usually grown to a maximal size (the tree-growing process stops when no further splits are possible because of the lack of data). Such a tree can easily lead to overfitting when minute details are learned at the expense of generalization.

The fully-grown tree can be then pruned back towards the root in order to remove splits that contribute least to the overall performance.

The CART mechanism includes (optional) automatic class balancing and automatic missing value handling, and allows for cost-sensitive learning (where different classes have different costs of misclassification) and probability estimation. CART can handle both continuous and categorical features.

CART is an unstable classifier whose decision splits and predictions can change due to small changes in the training data (this is called high variance). This is in the contrast to Naïve Bayes and $k$-nearest neighbors, considered earlier in this book, that are relatively stable with respect to small changes of the training data (it is said that these both classifiers have low variance). Due to its instability, CART is applied for bioinformatics tasks in classifier ensembles ((Diaz-Uriarte & Alvarez de Andrés, 2006), (Statnikov, Wang, & Aliferis, 2008)) rather than as a stand-alone classifier.

The major reason for instability of trees is the hierarchical nature of the process (Hastie, Tibshirani, & Friedman, 2003), leading to the effect of error propagation from the place it was made down to all the splits below it.

Hence, the practical usefulness of a single tree in making predictions, e.g., regarding cancer outcome or class, is rather limited (who needs an oracle that answers each time differently to the same question?). Instead, an ensemble of full-grown (i.e. non-pruned) classification trees provides better (both stable and more accu-
Computer-Aided Drug Design and Biological Evaluation of Novel Anti-Greek Goat Encephalitis Agents