Report(pruning)

Qi Ji

May 2020

1 Background knowledge

We can do pruning to optimize the tree size and reduce overfitting. Generally, there are two kinds of pruning: Pre-pruning and Post-pruning.

1.1 Pre-pruning

- Pre-pruning is also called early stoping.
- Pre-pruning stops the tree before it grows perfectly.
- People can set maximum depth of the tree to restrict the tree.

1.2 Post-pruning

1.2.1 REP(Reduced-Error Pruning)

REP is one of the simplest forms of Post-pruning.
First, calculate the error of a subtree $E_r(t)$. Then, calculate the error of each leaf node of this subtree $E_t(Tt)$. If

$$E_r(t) < \sum E_r(T_t)$$

Then, replace the subtree with a leaf node, whose label is determined by the majority label of the subtree. Repeat this process from the bottom of the tree to the top.

1.2.2 PEP(Pessimistic-Error Pruning)

For a leaf node with N samples and E errors, its error rate e is $\frac{E+0.5}{N}$. "0.5" is called penalty factor. For a subtree with L leaf nodes, its error rate is

$$p = \frac{\sum_{i=1}^{L} E_i + 0.5L}{\sum_{i=1}^{L} N_i}$$

Suppose that all the samples in a subtree is of binomial distribution $B(N, P)$, then the expectation and standard deviation of error before pruning are:

$$E_T = N * p = N * \frac{\sum_{i=1}^{L} E_i + 0.5L}{\sum_{i=1}^{L} N_i} = \sum_{i=1}^{L} E_i + 0.5L$$

$$\sigma = \sqrt{N * p * (1 - p)}$$

Expectation of error after pruning is:

$$E_t = N * e = N * \frac{E + 0.5}{N} = E + 0.5$$

If

$$E_t - E_T < \sigma$$

then prune the subtree. Repeat this process from the top of the tree to the bottom.

1.2.3 CCP(Cost-Complexity Pruning)

α is a real number called the complexity parameter, it's defined as follows:

$$\alpha = \frac{R(t) - R(T_t)}{|T_t| - 1},$$

where $R(t)$ is the error after pruning, $R(T_t)$ is the error of the original subtree and $|T_t|$ is the number of samples in the subtree.
Step1: Calculate the value of α for all the subtrees from the bottom to the top, each time prune the subtree with minimal α. Get a set $\{T_0, T_1, ..., T_M\}$, where T_0 is a complete tree and T_M is a root node.
Step2: Pick the best tree from $\{T_0, T_1, ..., T_M\}$, according to its performance on the testing sets.
*Note that if we use CCP to prune the tree, we should separate a testing set from the given training set before training begins.

2 Approaches to try

We adopted Pre-pruning by setting the maximum depth of the tree when building it. We considered three methods of Post-pruning: REP(Reduced-Error Pruning), PEP(Pessimistic-Error Pruning) and CCP(Cost-Complexity Pruning). Eventually we tried PEP and CCP, and CCP worked better. Therefore, we chose CCP to do the pruning.