Random forest

Random forest is a bagging method in integrated learning the data relationship as this kind of style

Fig.1

Aggreation [Boosting bagging(random forest)]

Fig.2

Decision Tree

Decision tree is a logically simple machine learning algorithm. It is a tree structure, that uses layered reasoning to achieve the final classification. The decision tree consists of the following elements:

Root node: contain the complete set of samples

Internal node : corresponding feature attribute test

Leaf node: represents the result of the decision.

Fig.3

Root node (leaf node , internal node(leaf node1,leaf node2))

In the prediction, a certain attribute value is to judge at the internal node of the tree and which branch node is entered according to the judgment result until the leaf node is reached .and classification result is obtained.

This is a supervised learning algorithm based on the if-then-else rules. These rules of the decision tree are obtained through training rather than artificially and it is the simplest machine learning algorithm. Easy to implement highly interpretable, and fully in line with human intuitive thinking.

When we decide apply this method we should mention decision tree first.

Decision tree is a very simple algorithm it is highly explanatory and conforms to human intuitive thinking ,this a sort of algorithm based on the if-then-else rule. The above picture can intuitively express the logic of the decision tree/

So we call the RF as random forest in the following text.

RF are made of man decision trees and there is no correlation between different decision trees. When perform the classification task, then new input sample enters and each decision tree will get its own classification result, and which classification result of the decision tree most ,then random forest will use this result as the final result.

Then there 4 steps to construct a random forest

1. random sampling training decision tree
2. randomly select attribute as node spilt attribute
3. repeat the step 2 until the data can’t split anymore
4. build large mount of decision tree and forming forest

A sample with a sample size of N is drawn N times with replacement , and one sample is drawn each time ,then finally N sample are formed. The selected N samples are used to train a decision tree as the samples at the root node of the decision tree

When each sample has M attribute when each node of the decision tree needs to be spilt, randomly select m attribute from M attributes and satisfy the condition m<<M the use a certain strategy(such as information gain)from these m attribute to select 1 attribute as the split attribute of the node

In the decision tree formation process, each node must be split according to the step 2,Until it can’t split again Note that no pruning is done during the formation of the entire decision tree.

According to the steps 1-3 a large number of decision trees are created, which constitutes a random forest.

But this method still have advantage as well as disadvantages as follows

Positive points including:

1. it can come out with very high dimensional(features )data and need to reduce dimension, it is not necessary to make feature selection
2. it can judge the importance of the feature
3. can judge the interaction of the feature
4. Not easy to overfit
5. 1Training speed is faster, easy to make parallel method
6. It is relatively simple to implement by the coding
7. For unbalance data sets, it balances the error
8. If a large part of the features are lost, accuracy can still be maintained.

But the negative affect is still markable enough to neglect

1. RF method have been shown to fit over certain noisy classification or regression problem.

2. For data with different values, attributes with more value will have a greater impact on random forest, so the attribute weight generated by random forest on such data are not credible

RF are commonly used machine learning algorithms that can be used for both classification and regression problems. And what’s more ,

Classification of discrete value, Regression of continuous values, Unsupervised learning clustering, and Abnormal point detection.