**AdaBoost with Decision Tree**

***Why should you care? –* AdaBoost** is a practical and easier boosting algorithm, which is an ensemble combining the weak classifiers produced by the weak learning algorithm into a single composite classifier. **Decision Tree (CART)** is a simple but powerful learning algorithm, which addresses the problems of tree structure design, feature selection, and decision rules to be used at each internal node. **AdaBoost with CART** is a more advanced algorithm to realize binary classification over the training data and predict the category of new input by the same rules and parameters. It is one of the important models in big data analysis (BDA) to construct the relationship between features and targets. This model improves the accuracy of prediction on whether a start-up succeeds or fails due to its high robustness to noise.



Figure 1: The relationship between AdaBoost and CART, and the framework of AdaBoost with CART.

***Approach –*** The main idea of AdaBoost with CART is focusing on training samples that the previous classifiers got wrong and allocating each weak classifier with appropriate weight. There are four key steps to implement the model: (1) At each iteration, **re-weight** the training samples by assigning larger weights to samples that were classified incorrectly. (2) Train a new base classifier based on the re-weighted samples. (3) Add it to the **ensemble of classifiers** with the right weight. (4) Repeat the process many times.

From the algorithm’s perspective, the specific process of the model is:



Figure 2: The mathematic algorithm of AdaBoost with CART.

From the programming’s perspective, the AdaBoost is implemented by functions or tools in the sklearn library.



Figure 3: The code that implements the AdaBoost with CART.

***Results –*** Among the current classification algorithms, **Random Forest** is the most appropriate benchmark for appraising the AdaBoost with CART. The distinction between these two algorithms is the way to select weak classifiers into the strong classifier. It is uncertain to determine which algorithm is better, and each model has a different performance and an emphasis according to given tasks. However, AdaBoost provides higher accuracy of prediction than Random Forest in forecasting whether a start-up succeeds or fails, with 85.714% accuracy.

|  |  |  |
| --- | --- | --- |
|  | **AdaBoost** | **Random Forest** |
| *Ensemble Method* | Boosting | Bagging or Bootstrap (random sampling) |
| *Feature Selection* | Necessary | Optional |
| *Performance* | Reduces bias | Reduces variance |
| *Suited in* | Flexible weak classifiers, Consideration of classifiers’ weights | Low noise of sample,  Quick learning from sample |

Table 1: Comparison between AdaBoost and Random Forest.

***Pros/Cons***

* Pro: Interpolation (perfect prediction in sample) was achieved after relatively few iterations.
* Pro: Generalization error continues to drop even after interpolation is achieved and maintained.
* Con: The later iterations lead to an “averaging effect”, which causes AdaBoost to behave like a random forest and overfits data.

***Conclusion –*** AdaBoost with CART is an efficient and precise ensemble classification algorithm because of its re-weighting contribution to each weak classifier. Compared with simple classifiers, it bases on a designed algorithm and provides robust predictions. Therefore, it is a reliable model to complete the binary classification.