A
REPORT
ON

K-Random Forests

By

Aarav Goel 2020A1PS1696P

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE,
PILANI (Rajasthan)

(March, 2023)

Introduction

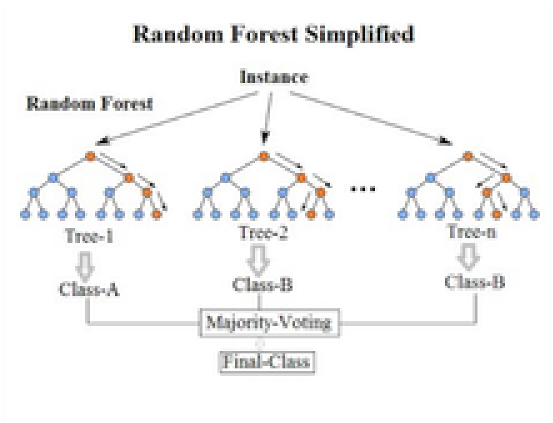
Random Forest is a well-known approach for supervised learning that combines the predictions of numerous decision trees to improve accuracy and reduce overfitting. K random forest is a variant of Random Forest that enhances the original method by providing a new parameter that determines the maximum number of characteristics utilised for each decision tree split.

How K Random Forest Works

The K random forest algorithm is comparable to the standard Random Forest algorithm, including the K parameter. This option determines the maximum features required to build each decision tree split. K random forest can reduce overfitting and improve model generalisation by restricting the number of features.

Below are the steps necessary to construct a K random forest model:

1. Choose a chunk of the training data at random.
2. Choose K characteristics at random from the sample of data.
3. Create a decision tree based on the specified features.
4. Repetition of steps 1-3 will result in a forest of decision trees.
5. Combine the predictions of all the trees in the forest to generate a prediction.



Advantages of K Random Forest

The K random forest method has some advantages over Random Forest:

1. Reduced overfitting: By restricting the number of features utilised for each split, K random forest can prevent overfitting and enhance the model's generalisation.
2. By picking the optimal subset of features for each split, K random forest can enhance its accuracy.
3. Using a smaller collection of characteristics, K random forest can train more quickly than the standard Random Forest algorithm.

Dataset Description

Parkinsons Dataset:

Parkinson's dataset is a collection of biomedical data from Parkinson's disease patients. This dataset is frequently used in machine learning and data analysis to construct predictive disease diagnosis and progression-tracking models.

Each of the 195 records in the collection represents a patient with Parkinson's disease. Each patient's voice was recorded 15 times, and each recording was evaluated for numerous speech signals and characteristics, such as jitter, shimmer, and other acoustic characteristics.

This dataset's target variable is a binary indicator of whether or not the patient has Parkinson's disease. In addition to the demographic and clinical characteristics of the patients, such as age, gender, and time since diagnosis, the dataset contains additional demographic and clinical information.

The Parkinson's dataset can be utilised for various machine-learning applications, including classification, clustering, and regression. The dataset enables researchers to investigate the association between speech signal features and the course of Parkinson's disease. It can also aid in identifying new biomarkers that may improve the accuracy of disease diagnosis and disease progression tracking.

Iris Dataset:

The Iris dataset is a commonly used dataset in machine learning and data analysis, comprising information about the physical characteristics of many Iris flower species. Frequently, the dataset serves as a benchmark for classification models and exploratory data analysis.

The collection contains 150 instances, each of which represents a single flower. Each instance describes four characteristics of the flower: the length and breadth of the sepals and the length and width of the petals. This property is measured in millimetres.

The Iris species represented in the dataset are Iris setosa, Iris versicolor, and Iris virginica. This dataset's objective variable is the species of each flower, a categorical variable.

The Iris dataset is frequently used for exploratory data analysis since it is tiny and straightforward to comprehend. As a well-understood and standardised dataset with a defined goal variable, it is frequently utilised as a benchmark for classification methods.

Overall, the Iris dataset is a valuable resource for investigating correlations between flower qualities and species classification and for developing predicting algorithms to categorise new Iris flower examples based on their characteristics.

Wine Quality Dataset:

The Wine Quality dataset is a common dataset in machine learning and data analysis, providing data on the physicochemical properties and quality ratings of red and white wines.

The dataset comprises a total of 12 features, such as pH, alcohol concentration, residual sugar, and citric acid. These characteristics are used to forecast the wine's quality on a scale from 0 to 10. The dataset contains red and white wines, with 1599 red and 4898 white wine instances, respectively.

In addition to chemical characteristics, the dataset contains categorical factors such as the type of wine (red or white) and its quality rating. The sensory data-based quality assessment goes from 0 (extremely poor) to 10 (outstanding) (very excellent).

The Wine Quality dataset is frequently used for regression analysis and classification tasks, and it is a valuable tool for investigating the links between wine qualities and quality ratings. Based on its chemical features, it can also be utilised to construct predictive models for estimating wine quality.

Auto-mpg dataset:

The Auto-MPG dataset is widely used in machine learning and data analysis, comprising information about the fuel consumption and other characteristics of automobiles from the 1970s and 1980s. Regression models and exploratory data analysis frequently utilise the dataset as a benchmark.

The collection consists of 398 instances, each representing an automobile model. Each instance includes the car's horsepower, weight, displacement, and acceleration. This dataset's goal variable is each vehicle's miles per gallon (MPG) rating, representing fuel efficiency.

The Auto-MPG dataset also includes other categorical factors, such as the model year and the car's country of origin (American, European, or Japanese) (ranging from 1970 to 1982). These variables can be used to investigate patterns and variances in the data across various car groupings.

Overall, the Auto-MPG dataset is a valuable resource for investigating correlations between car qualities and fuel efficiency and for developing prediction models to estimate MPG values based on car parameters.

Results

For the Parkinson's dataset, the patient's status, i.e., whether they have Parkinson's or not, was utilised as the goal variable. A random forest was trained using Holdout, Cross Validation, and Bootstrapping techniques. The accuracy of the Holdout method was interestingly determined to be 77.97%, whilst Bootstrapping and Cross-validation accuracy was 100%.

For the Iris Dataset, the species attribute was utilised as the target variable, and the accuracy of the holdout approach was determined to be 97.78%, whilst the accuracy of Cross-Validation and Bootstrapping was 100%.

For the Wine quality data collection, the target attribute was the taste attribute (if the quality is greater than 5, then the taste is good; otherwise, the taste is bad). With quality = 5,6, the flavour is normal. For the holdout approach, the accuracy was 83.79 per cent, and for the iris and Parkinson's datasets, the bootstrapping and cross-validation accuracy was one hundred per cent.

The nation of origin was used as the goal variable for training random forests on the auto-mpg dataset (1 - USA, 2-Europe,3-Japan). The holdout accuracy was 2.5%, cross-validation accuracy was 5%, and bootstrap accuracy was 11.06;%.

Discussion

The results that were obtained on the different datasets while training random forest using different methods for data splitting were similar to those in the research paper **K-Random Forests: a K-means style algorithm for Random Forest clustering.** The accuracy of the random forest algorithm was pretty good for the iris, parkinsons and wine quality datasets and even achieved an accuracy of 100% for cross validation and bootstrapping. But it was low for the auto-mpg dataset, with the bootstrap accuracy being the best of the three methods at 11.06%.

Method	**Parkinson's**	**Iris**	**Wine quality**	**Auto-mpg**
Holdout-Method	77.97%	97.78%	83.79%	2.50%
Cross Validation	100%	100%	100%	5%
Bootstrapping	100%	100%	100%	11.06%

Obtaining a 100% accuracy for some of the above datasets may indicate overfitting. This might be due to several reasons:

1. There are too many features/factors in the datasets that are being taken into account while training the model.
2. The model has been trained for a large number of iterations and it has memorised the data instead of finding out the underlying pattern in it.

If the number of trees were reduced in the iris dataset in the range of 3-8, the accuracy is 98.67%. Similarly, if the number of trees were reduced to 3-4 in the Parkinson's dataset, the accuracy obtained was 98.98%.

Conclusion

K random forest is a potent variant of the Random Forest algorithm that can enhance precision and prevent overfitting. It accomplishes this by restricting the number of characteristics utilised for each decision tree split. Despite its limitations, K random forest is a powerful tool for machine learning practitioners and has a wide range of applications. And this algorithm was tested on 4 datasets which gave good results.