A Comparative Study of Random Forest & K – Nearest Neighbors on HAR dataset Using Caret

Kella BhanuJyothi¹, K Hima Bindu² and D. Suryanarayana³
¹PG Scholar, CSE, Vishnu Institute of Technology, Bhimavaram, A.P, India
²,³Professor, CSE, Vishnu Institute of Technology, Bhimavaram, A.P, India

Abstract—In this paper, we have compared the accuracy results of the Classification models Random Forest Algorithm and K – Nearest Neighbors on UCI HAR (Human Activity Recognition) dataset and selecting the best classifier among the models to test the dataset. The best classifier is choosing by accuracy of the model. The classifiers for the dataset are built from the methods ‘rf’, ‘parRF’ and ‘knn’ by using Caret package in R. The classifier is used to build on the data which have high quality. If, the data contains incorrect and noise information then the accuracy we will receive is reduced. So, before preparing the classifier the dataset should be preprocessed to get exact accuracy. From the three methods, we found that the method ‘rf’ have high accuracy 93.13% on the trained model Hence, which is considered as a best classifier and applied the method ‘rf’ on the test data of HAR dataset.

Index Terms—Classifier, rf, ParRF, knn, Accuracy, HAR Dataset, caret package.

I. INTRODUCTION

Consider audio and video data, social media posts, 3D data or geospatial data, this kind of data is not easily categorized or organized [1]. To overcome this challenge, a range of automatic methods for extracting useful information has been developed, among them one of the techniques is classification. Classification is used to assigns items in a collection to target categories or classes [2]. The aim of the classification is to accurately predict the target class for each case in the data. Decision Tree is one of the most popular classification techniques [3] and it is a tree-structured plan of a set of attributes to test in order to predict output.

Terabytes of information is generating day by day in various formats and it is increasing day to day but the data is often incomplete, inconsistent and/or lacking in certain behaviours or trends, and is likely to contain many errors, which may lead to inaccuracy in the data [4]. To resolve this issue, Data Pre-processing method is used. Data Pre-processing is a data mining technique that involves transforming the raw data into understandable format [5].

Data Splitting segregates the data into two parts. One part of the data is used to develop a predictive model and the other is to evaluate the model’s performance [6]. Random forests are an ensemble learning method for classification and regression that operate by constructing many decision trees at training time and outputting the class that is the mode of the classes output by individual trees.

HAR dataset stands for human activity recognition dataset, which is a collection of 6 different activities – laying, walking, sitting, standing, walking_upstairs, walking_downstairs of a human being monitored by some social workers. This dataset come across different stages in preparation [7].

II. RELATED WORK

The HAR dataset is prepared from the sensor signals which have been taken from the smartphone Samsung Galaxy S II attached to the group of 30 volunteers within an age bracket of 19-48 years [8]. The sensor signals are considered for the 6 different activities bee observed by the 30 volunteers. The data is of 561 features and 10299 instances. The accuracy calculation for this large dataset of 561 attributes is quite typical and the accuracy we get is not the correct one [11]. In order to get proper accuracy pre-processing of the data is required.

The caret package stands for Classification And Regression Training is a set of functions that attempt to streamline the process for creating predictive models [8]. Below are the tools which are included in the package -

- Data splitting
- Pre-processing
- Feature selection
Model tuning using resampling
• Variable importance estimation

One of the main reasons to use cross-validation instead of conventional validation (e.g. the dataset is divided into two sets one is 70% for training and the other part is for 30% for test) is that there is not enough data available to divide it into separate training and test sets without losing significant modelling or testing capability [6]. In these cases, a better way to properly estimate model prediction performance is to use cross-validation. Cross-validation combines measures of fit (prediction error) to derive a more accurate estimate of model prediction performance.

Accuracy is how often the model trained is correct, which is depicted by using Confusion Matrix. A confusion matrix is the summary of prediction results on a classification problem [7]. The number of correct and incorrect predictions are summarized with count values and separated by each class, which is the key to the confusion matrix.

III. WORKING METHODOLOGY

The pre-processing techniques we have used are Zero and Near Zero Variance Predictors. Zero Predictors removes the predictors which have one unique value across samples [4]. Near Zero Variance is using to remove the predictors which are unique values relative to the number of samples and big ratio of the frequency of the most common value to the frequency of the second most common value.

Next, removing the correlated predictors from the dataset to get the quality data, the generated dataset contains 277 variables out of 561 attributes. K-fold cross validation is applying on the train dataset which is preprocessed to get more accuracy in predictive model [4]. The K value chosen is 7 here.

The accuracy of the HAR dataset is predicted for 3 methods, those are – rf, parRF & knn. The models are trained by using the 3 methods - rf, ParRF & knn and we are determining the best model suited for the HAR dataset among the predicted models by accuracy of the model.

A. Method – rf

Method we have used is “rf” stands for Random Forest to predict the model on HAR dataset. Random Forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and producing the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees [9].

The Random Forest model is creating for the HAR dataset by using train() function in caret package of R [8]. The accuracy derived from the ‘rf’ method is 93.13%

B. Method – parRF

The method parRF is short for Parallel Random Forest. By Parallel Processing computation time is decreased, and it is configured by using the method make PSOCK cluster(n) in windows where “n” number of cores using as worker threads for the parallel computation [9].

The accuracy for the randomly selected predictors by ParRF method is 92.98%

C. Method – knn

K - Nearest neighbors algorithm is a simple one that stores all available cases and classifies new cases by a
majority vote of its k neighbors. This algorithm separates unlabelled data points into well-defined groups [10].

![Figure 3 Accuracy Graph of Knn](image)

The accuracy graph of knn is denoting that for K value 15, the accuracy meets above 86% and it is the high value got by this knn method using train() function in R.

The accuracy derived from the knn method is 86.28%.

IV. RESULTS

The best classifier is determining from the accuracy of the model. Therefore the classifier is chosen among these three rf, parRF & knn methods is ‘rf’ as it is have high accuracy i.e. 93.13%.

The accuracy is determined from the 3 methods rf, parRf, knn are determined in confusion matrix as below

<table>
<thead>
<tr>
<th>Method</th>
<th>rf</th>
<th>parRF</th>
<th>knn</th>
</tr>
</thead>
<tbody>
<tr>
<td>Activity</td>
<td>Walking</td>
<td>Upstairs</td>
<td>Downstairs</td>
</tr>
<tr>
<td></td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>0.92</td>
<td>0.92</td>
<td>0.92</td>
</tr>
<tr>
<td>Walking</td>
<td></td>
<td>0.91</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>0.92</td>
<td>0.92</td>
<td>0.92</td>
</tr>
<tr>
<td>Upstairs</td>
<td></td>
<td>0.91</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>0.92</td>
<td>0.92</td>
<td>0.92</td>
</tr>
</tbody>
</table>

![Table 1 Comparison Accuracies of rf, parRF, knn methods](image)

Before to apply ‘rf’ method on test dataset, the dataset is to be preprocessed as like train dataset and then apply ‘rf’ method on HAR test dataset, the accuracy received is 94.88%.

V. CONCLUSION

This paper works on to determine the best classifier from the training methods rf, parRF & knn. The accuracy of the dataset is depends on how much quality is the data and by the method which is trained. The preprocessing technique and splitting of the data techniques used are perfect for the HAR dataset which results to the good accuracy value.

REFERENCES


