COSC 5557
Auto-Sklearn Fail

Almountassir Bellah Aljazwe

May 2024

1 Sequence of Changes

The following is the process of changes undertaken to improve the automated machine learning results obtained by the "AutoSklearn" package. These changes build on top of each other in the listed order. The random state is kept similar to the initial value (42) and the default time budget (5 minutes) is also not modified at all.

1.1 Modification of Splitting Strategy

Initially, the provided code used the "sklearn.model_selection.train_test_split" module to split the data into training and testing splits. My initial thought was to replace the use of that module to an alternative module that preserves the ratio of the target feature labels from the initial dataset.

Here is a comparison of the test target-data split from the default "train_test_split" module and the alternate "StratifiedShuffleSplit" module :

Table 1: Whole Dataset Distribution

3	4	5	6	7	8
10/1559	53/1559	681/1559	638/1559	199/1559	18/1559
$\approx 1\%$	$\approx 3\%$	$\approx 43\%$	$\approx 40\%$	$\approx 12\%$	$\approx 1\%$

Table 2: Train-Test-Split Test Target Distribution

3	4	5	6	7	8
1/1559	13/1559	164/1559	169/1559	48/1559	5/1559
$\approx 0\%$	$\approx 3\%$	$\approx 41\%$	$\approx 42\%$	$\approx 12\%$	$\approx 1\%$

Table 3: Stratified-Shuffle-Split Test Target Distribution

3	4	5	6	7	8
2/1559	13/1559	170/1559	160/1559	50/1559	5/1559
$\approx 0\%$	$\approx 3\%$	$\approx 42\%$	$\approx 40\%$	$\approx 12\%$	$\approx 1\%$

As can be seen, there is no real difference and as a result, it is not a surprise that the accuracy score, of the Auto Sklearn classifier, still was lower than the default Random Forest classifier. The Auto Sklearn classifier achieved an accuracy score of 0.645 while the Random Forest classifier achieved an accuracy score of 0.655.

1.2 Modifying the Default Resampling Strategy

When initializing an instance of an "AutoSklearnClassifier" object, you can view the default parameters using the "get_params()" function. The results of this show the default resampling strategy of "holdout" which is a 67:33 (train:test) split, according to the documentation. Instead of the default "holdout" strategy, I went with a 10-fold cross validation resampling strategy, on a test split size of 0.8, in an attempt to introduce less bias and variance with the model evaluations.

This, however, had no apparent effect as the accuracy performance with AutoSklearn (0.6325) still was worse than the default Random Forest Classifier (0.655).

1.3 Modifying the Scoring Metric

After running the "AutoSklearnClassifier" object, you can check some statistics that summarize the classification results. An important statistic is the scoring metric used in the automated process; the default scoring metric used from the "AutoSklearnClassifier" object is accuracy; this is not the optimal scoring metric for an imbalanced dataset such as the "Wine Quality" dataset in the file. As such, the accuracy scoring metric will be modified to the "f1_weighted" metric to better consider the dataset imbalance. This time, there is a relatively slight improvement in performance, with the "AutoSklearnClassifier" resulting in an accuracy score of 0.6725 and the Random Forest Classifier resulting in an accuracy score of 0.655. Although an improvement, this may simply be due to random chance, so more improvements may be required to make sure.