PROJECT REPORT: C S 487/519 Applied Machine Learning I

Project 8: Ensemble approaches

Written by

Omid Jafari

Department of Computer Science

New Mexico State University

November 2018

Preprocessing and notes:

Mammographic dataset contained some missing values; therefore, we used the SimpleImputer function from Sklearn library to replace the missing values with the mean of each attribute.

After that, as usual, we loaded the datasets and standardized them.

Splitting to training and testing sets were done on a 0.7/0.3 basis.

Decision Tree was chosen as the base classifier.

Ensemble methods were executed using several different parameters and compared to the base classifier. Since parameters “criterion”, “random_state”, and “max_depth” were common between the base classifier and ensemble algorithms, we did not consider changing them. The only parameters that made sense to tune them were “bootstrap” for Bagging algorithm and “learning_rate” for AdaBoost algorithm.

Results:

Running time of Digits dataset (ms)		
Decision Tree (Base)	Training	14
	Prediction of train set	0
	Prediction of test set	0
Random Forest (Default params)	Training	47
	Prediction of train set	6
	Prediction of test set	4
Bagging (Default params)	Training	206
	Prediction of train set	12
	Prediction of test set	6
AdaBoost (Default params)	Training	14
	Prediction of train set	2
	Prediction of test set	1
Bagging (bootstrap=False)	Training	352
	Prediction of train set	13
	Prediction of test set	6
Bagging (bootstrap_features=True)	Training	212
	Prediction of train set	12
	Prediction of test set	6
AdaBoost (learning_rate=0.5)	Training	15
	Prediction of train set	1
	Prediction of test set	1
AdaBoost (learning_rate=2.0)	Training	14
	Prediction of train set	1
	Prediction of test set	1

Accuracy of Digits dataset (%)		
Decision Tree (Base)	Train set	1
	Test set	0.861
Random Forest (Default params)	Train set	1
	Test set	0.970
Bagging (Default params)	Train set	1
	Test set	0.946
AdaBoost (Default params)	Train set	1
	Test set	0.848
Bagging (bootstrap=False)	Train set	1
	Test set	0.869
Bagging (bootstrap_features=True)	Train set	1
	Test set	0.970
AdaBoost (learning_rate=0.5)	Train set	1
	Test set	0.848
AdaBoost (learning_rate=2.0)	Train set	1
	Test set	0.848

Running time of Mammographic dataset (ms)		
Decision Tree (Base)	Training	1
	Prediction of train set	1
	Prediction of test set	0
Random Forest (Default params)	Training	21
	Prediction of train set	4
	Prediction of test set	2
Bagging (Default params)	Training	25
	Prediction of train set	4
	Prediction of test set	3
AdaBoost (Default params)	Training	37
	Prediction of train set	5
	Prediction of test set	3
Bagging (bootstrap=False)	Training	32
	Prediction of train set	4
	Prediction of test set	3
Bagging (bootstrap_features=True)	Training	23
	Prediction of train set	3
	Prediction of test set	3
AdaBoost (learning_rate=0.5)	Training	38
	Prediction of train set	5
	Prediction of test set	3
AdaBoost (learning_rate=2.0)	Training	36
	Prediction of train set	5
	Prediction of test set	3

Accuracy of Mammographic dataset (%)		
Decision Tree (Base)	Train set	0.955
	Test set	0.727
Random Forest (Default params)	Train set	0.952
	Test set	0.761
Bagging (Default params)	Train set	0.951
	Test set	0.765
AdaBoost (Default params)	Train set	0.955
	Test set	0.734
Bagging (bootstrap=False)	Train set	0.955
	Test set	0.723
Bagging (bootstrap_features=True)	Train set	0.932
	Test set	0.817
AdaBoost (learning_rate=0.5)	Train set	0.955
	Test set	0.723
AdaBoost (learning_rate=2.0)	Train set	0.955
	Test set	0.713

Analysis:

Random Forest and Bagging algorithms have the most improvement in accuracy. However, the running time of them has increased too.

Changing the learning rate of AdaBoost did not improve the accuracy and running time. On the other hand, using bootstrapping for the features in Bagging method improved accuracy.