COMPSCI 361 ASSIGNMENT 2

Ensemble algorithms

Haozhou SHEN, 731260510, hshe507

March 29, 2019

Dataset	(1) trees.RandomFo	(2) meta.AdaBoo	(3) meta.Baggin
iris	(100) 94.67(5.01)	95.40(5.74)	94.47(5.69)
car	(100) 94.67(1.74)	70.02(0.16) *	91.66(2.13) *
balance-scale	(100) 81.48(3.62)	71.77(4.24) *	83.68(3.65)
	(v/ /*)	(0/1/2)	(0/2/1)

Data and Algorithm analysis

For the first iris dataset, all three algorithms demonstrate the similar performance. However, the AdaBoostM1 algorithm with DecisionStump classifier could be considered as the best algorithms. AdaBoostM1(Boosting) has the best overall accuracy [U+FF08] 95.40>94.97>94.67 [U+FF09] and only a slightly higher deviation compared to other two algorithms.

For the car dataset. The algorithm with the best performance is the RanfomForest algorithm. The accuracy of RandomForest algorithm on car dataset is significantly higher than the outcome of AdaBoost and Bagging algorithms. From the observation to the raw car data set, the dataset has some noise attributes and only some of attributes can be considered usable. This is the reason why AdaBoost with DecisionStump has significantly worse result than the other two as DecisionStump starts at 1 attribute node each time. Also Bagging with REPTree takes all attributes into modelling so it will be influenced by the noise in the dataset. Generally, RandomForest performed best among the tests.

As to the last balance-scale dataset, I believe the Bagging algorithms with REPTree is the best choice for classification. Bagging with REPTree and RandomForest both generate results with nearly 82% accuracy and share similar variations. The performance of AdaBoost is much worse than the other two. Because the Bagging has a similar level of variation with RandomForest (3.62 and 3.65) and higher accuracy (83.68% $>$ 81.48%), it is believed to be the best algorithm for the balance-scale dataset.

Best algorithm

Overall, the RandomForest algorithm is considered to be most reliable. In the experiment, results generated from RandomForest algorithm has the best overall performance. In the 6 comparisons performed between RandomForest and another two algorithms, RandomForest provided similar accuracy level in 3 comparisons of 6 ones and outperformed Bagging and AdaBoost in another 3. Additionally, RandomForest algorithm can generate results with smallest variance in the most cases ($2/3 = 66.6\%$ in our experiment), which means RandomForest provides more stable and reliable output on different datasets. Finally, natural difference between three algorithms can lead to the different performances. As DecisionStump only starts at 1 attribute each time so it will significantly increase the error if choosing the wrong starting attribute, so eventually it will depreciate the performance of the AdaBoost algorithm. On another hand, bagging with REPTree will use all attributes when starting a classification, so the resulting bagging algorithm will be very sensitive to the noise in the raw data. This could generate very low accuracy when encountering complex datasets with strong noise. However, RandomForest algorithm only takes several attributes when start building a single tree instead of using all attributes, and bagging after each single decision trees can improve the accuracy with the resampling instances process. All these properties give RandomForest algorithm ability to have a better overall performance with different datasets.