CEE 412 Assignment 4

Houhao Liang

University of Illinois at Urbana-Champaign

12/10/2018

Summary

The overall accuracy of the test dataset for each dataset can be found in Figure as below. I have successfully passed first three datasets, however failed to implement the fourth dataset.

Balance						
	Decision Tree			Random Forest		
Accuracy	64.44%			72%		
Class	1	2	3	1	2	3
F-1 Score	0	0.748815	0.676923	0	0.753488	0.75

Led				
	Decision Tree		Random Forest	
Accuracy	85.89%		86%	
Class	1	2	1	2
F-1 Score	0.772727	0.897698	0.770774	0.898089

Nursery										
	Decision Tree					Random Forest				
Accuracy	97.45					98%				
Class	1	2	3	4	5	1	2	3	4	5
F-1 Score	0.96264	0.7399	0.983	1	0	0.9664	0.79	0.9811	1	0

Synthetic								
	Decision Tree				Random Forest			
Accuracy	24.70%				28%			
Class	1	2	3	4	1	2	3	4
F-1 Score	0.2475	0.2778	0.257	0.1979	0.26	0.2897	0.25	0.1849

Introduction

This project adopts random forest and decision tree method to make prediction. Gini Index has been used to select and linear combination method for random forest.

Decision Tree

The format of dataset for this project is LIBSVM format. In my decision tree, node represents a attribute and children are the value of this attribute. Tree class is used to build up the decision tree. Gini index was implemented as the attribute selection measure. The attribute is selected corresponding to the least Gini index.

Random Forest

Random linear combination of multiple decision tree has been implemented in my program. Each decision tree was trained by randomly selected dataset, and Gini index was calculated correspondingly. Finally, the prediction labels were selected by majority voting.

Parameters settings

Decision Tree

For my decision tree method, I haven't set a constant max depth. Considering all dataset can pass the time limit and overall accuracy, it's acceptable that the decision tree will not be stopped in advance.

Random Forest

The tree number of my forest is 100, and training dataset will choose data with a probability of 0.6. I have tried either change tree number or change the probability. After several manual iteration, 100 and 0.6 is acceptable considering the running time, overall accuracy, F-1 score, etc.

For the number of attributes, I set the number is the original number attributes divide by 1.2 and attributes are randomly selected. Therefore, for decision tree, the attributes are significant different, and it will only use about part of attributes to calculate Gini index. Thus, the result is more reasonable.

Conclusion

Overall

Basically, ensemble classification method promotes the overall accuracy, especially for balance dataset. But for Led and Nursery, the prediction accuracy of decision tree is high enough. It would be more efficient to adopt decision tree since less running time.

Balance:

The overall accuracy increases significantly by implementing random forest, from 62.44% to 72%. Particularly for class3, the F-1 score promotes about 8%. Random forest does very well in predicting class 3. Correspondingly, the overall accuracy raises.

Led & Nursery

Considering decision tree has achieved a very high accuracy, random forest actually promotes the overall accuracy a little bit. Probably for these two datasets, decision tree has the same prediction accuracy as random forest while it is more efficient with less running time.

Synthetic

For this dataset, I'm really sorry about this result. I tried a lot of method to debug my method. Considering the first index in this dataset start from 0, while others start from 1, that would be a issue in my code. But after I fixed the issue in my classifier, it's still inaccurate. For future work, I would continue debug my code for this dataset.