COMPSCI 361 ASSIGNMENT 1

Real data set identification

Haozhou SHEN, 731260510, hshe507

March 15, 2019

Introduction

After the analysis of the decision tree algorithm results, the obscured2 data set is believed to be the real file whose class labels was not modified.

Methods used for analysis

Comparison of accuracy

First way to determine the signal is comparing two accuracy with the percentage of majority class in the data set. Even though one of the data set was randomly shuffled. The records with the class label C take up $33950/38160 = 88.97\%$ in both data sets. This makes C the majority class in the data. However, when using J48 decision tree algorithm to classify the data, obscure1 gets an overall accuracy of 88.9675% and obscured2 gets a accuracy of 92.2327%. Comparing these two results with the original proportion of the majority class, the first result from obscured1 is in the same level while the obscured2's result is significantly higher than the majority proportion. This underlines that there is some deeper correlation between attributes in the second data set rather than just using majority class for classification in the first data set. So the obscured2 is the data set which more likely to have a signal.

Generated trees comparison

Another method used to determine the real data set is using the decision trees generated by the algorithm. First decision tree has only one node which means it will classify all records to the label C. However, the second decision generated from obscured2 data set has a height of 4 and overall 1623 leaves, which implies there are additional information provided by other attributes. This can also be checked with direct visualization of the full_name and country_name attribute in the second data set. Thus, we can conclude that the obscured2 data set has a real signal.

(a) obscured1 decision tree

(b) obscured2 decision tree

Figure 1: Generated decision trees comparison

Conclusion

Comparing two methods used for identification, the second method is explicitly easier to use. Even through the difference of two result in first method can help us distinguish between two data sets, it is possible to say that sometimes higher accuracy can be achieved randomly by luck. However the decision tree diagram of obscured2 data set, which was shown in the method 2, presents a clear and strong logic relations between attributes and classes during the classification. Additionally, the diagram method is a more obvious and intuitive way to use because it can give us an overall view of classification process. Thus, I suggest the generated trees comparison method is better.