COMPSCI 361 Assignment 3

Haozhou SHEN 731260510

The assignment is using a Naïve Bayes classifier to predict the label of giving data evaluated under a 30% holdout validation set. Preprocessing process was conducted on raw data set with Weka Explorer and the comparison of performances of different pre-processed data sets was conducted using Weka Experimenter.

Preprocessing Techniques.

Overall performances:

图片包含 屏幕截图

已生成极高可信度的说明

Replace missing value with average value:

For the giving data set, replacing missing field with average value will give us a higher accuracy. After testing via skit-learn, if we eliminate tuples with missing values then all instances will be deleted. On the other hand, if we use a global value (e.g. -10000 in experiment), this could increase difficulty to find a signal. Thus using average to fill in missing fields will give us higher accuracies and therefore improve performance.

Feature selection:

As the giving data set is highly dense in attributes but only few instance records, feature selection can be used to minimize the bias and improve the performance. Shown by the table above, after feature selection which eliminate relatively irrelevant and redundant attributes from the raw data. Performance of Naïve Bayes algorithm was significantly raised.

Best attributes:

Attributes selection process was conducted using featureSelection filter with CorrelationAttributeEval evaluator and Ranker search. The Best 5 attributes chosen from the processed data (replace average + remove extreme values + feature selection) are shown in table below:

图片包含 屏幕截图

已生成极高可信度的说明