COMPSCI 361 ASSIGNMENT 3

Preprocessing techniques

Haozhou SHEN, 731260510, hshe507

April 15, 2019

Pre-processing techniques

Preprocessing done	Accuracy with split validation (10 repetitions)/Std. Deviations
Raw data	65.73(2.55)
Replace missing with global constant	61.89(4.57)
Replace missing with average value	66.10(2.53)
Clean noisy data with binning	65.06(3.39)
Clean noisy data with removing extreme values	63.07(4.01)
Reduce Feature selection	67.12(1.79)
Binning + Feature selection	70.68(1.45)
Replace average + Feature Selection	67.33(1.69)
Replace average + Remove extreme value + Feature selection	81.96(4.46)

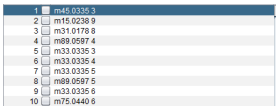
Replace missing value with average value:

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 accuracy and therefore improve performance.

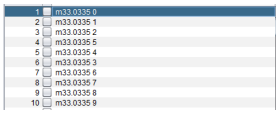
Reduce dimensionality with feature selection

As the giving data set has highly dense attributes but have 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



(a) Attributes sorted by correlation on processed data



(b) Attributes sorted by correlation on raw data

Attributes selection process was conducted using featureSelection filter with CorrelationAttributeEval evaluator and Ranker search. It is obvious to see that best attributes given by filter in raw data sets are highly inter-connective, which implies there are many redundant attributes. After pre-processing, we can find out that now attributes which have higher correlation to the target class all shows different pattern and are more independent to each other (can be shown via explorer). Thus these attributes are considered best attributes for classification.