Compsci 361 Assignment 4

Hasnain Cheena
190411106
hche737

After examining the dataset, the two most notable aspects are that portions of data are missing and there are far more attributes than instances. Therefore, data imputation to replace missing values followed by feature selection to remove irrelevant attributes are key stages of my pre-processing pipeline.

Other key things to note are the metric used to examine performance was accuracy. Accuracy was calculated through 10-fold cross validation with 30% holdout. The results were averaged to get a reliable value for accuracy. Moreover, to evaluate performance improvement comparison to a baseline was necessary. Therefore, a majority class classifier was used which had an accuracy (taken as a baseline) of 65.3%.
The train-test split was completed before any imputation and feature selection to ensure that information was not shared from the test set to the training set. Furthermore, the Naïve Bayes classifier was then run after each pre-processing action to determine the increase in performance.

Pre-processing Approach

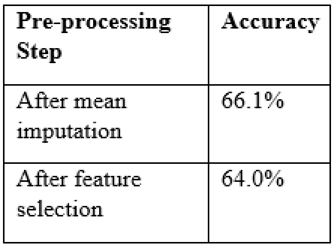
First data imputation was performed. Two forms of imputation were tested against one another; mean imputation and class-mean imputation. In mean imputation the missing values are replaced with the attribute mean. In contrast, in class-mean imputation missing values are subset by class and replaced with the attribute mean of the class. Table 1 and Table 2 below show that mean imputation and class-mean imputation have similar performance and are only marginally better than baseline. This may be because there are still many irrelevant features within the dataset, making it difficult for the classifier to pick up a signal.

Therefore, after imputation, feature selection using ReliefF was performed on both sets of imputed attributes. Table 2 shows that feature selection on the class-mean imputed features greatly increased performance relative to the mean imputed features (shown in Table 1). This proves that the irrelevant features were deterring performance (in the case of the class-mean imputed features) making it hard for the classifier to capture the underlying pattern in the data. Furthermore, the low accuracy when using the mean-imputed features indicates that the attribute mean is different between the classes and thus when using the same attribute mean for both classes the signal is disrupted.

Top 5 Feature Selection

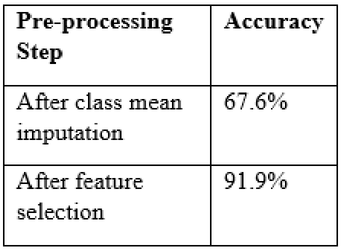
ReliefF calculates a weight for each feature which corresponds to the importance of the feature. Therefore, the ReliefF weight vector was used to determine the most important features. Using this weight vector, the features were ranked across each cross-validation run. The top 5 features within each run were extracted. The counts of these features were taken and placed within a ranking list. This list (shown in Table 3) was used to find the overall top 5 features. This method was necessary because each time ReliefF is performed it produces a different ranking order of the features as a result of the randomised nature of the train test split. Therefore, by creating a ranking list across all the cross-validation runs, features that are truly important should appear frequently.

The top 5 features are: 719, 2669, 699, 3019 and 309.



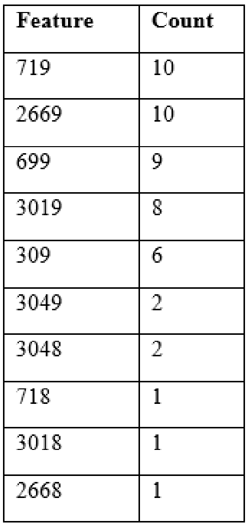
Pre-processing Step	Accuracy
After mean imputation	66.1%
After feature selection	64.0%

Table 1: Mean Imputation and ReliefF



Pre-processing Step	Accuracy
After class mean imputation	67.6%
After feature selection	91.9%

Table 2: Class-Mean Imputation and ReliefF



Feature	Count
719	10
2669	10
699	9
3019	8
309	6
3049	2
3048	2
718	1
3018	1
2668	1

Table 3: Feature Ranking List