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 attributes than instances. Therefore, data imputation to replace missing data and feature selection to remove irrelevant features are required on this dataset. These two elements are the motivations for my pre-processing pipeline.

Other key things to note are the metric used to examine performance was accuracy. Furthermore 10-fold cross validation was run with 30% holdout and results averaged to get a reliable value for accuracy. Moreover, to evaluate improvement in performance, as a baseline for comparison a majority class classifier was used. It had an accuracy 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 during any pre-processing. The Naïve Bayes classifier was then run after each pre-processing action to determine increase in performance.

*Pre-processing Approach*

First data imputation was performed. Two forms of imputation were tested against one another; overall mean imputation and class mean imputation. In overall mean imputation the missing values are replaced with the attribute mean. In contrast in class mean imputation missing values within a class are replaced with the attribute mean of that class. The table below shows that class and mean imputation are relatively similar and only marginally better than baseline performance when performed on their own (66.1% versus 67.6%). This is because there are still so many irrelevant features which are making it hard for the classifier to pick up a signal.

Therefore, after imputation, feature selection using ReliefF was performed on both sets of imputed features. From the table below, it can be observed that feature selection on the class mean imputed features is increased performance significantly more than the overall mean imputed features. This indicates that the attribute mean is different between the classes and thus the when using the same attribute mean for both classes the signal is disrupted. Further it proves that the irrelevant features were deterring performance by making it hard for the classifier to pick up the signal.

*Top 5 Feature Selection*

ReliefF calculates a weight per feature which can be used to select important features. Using this weight vector, the features were ranked across each cross-validation run. The top 5 features within each run were extracted and then placed within a ranking list. This list (shown in table 3) was used to find the overall top 5 features. This method was used because due to the randomised nature of the train test split each time ReliefF (and therefore each cross-validation run) is performed it produces a different ranking order of the features. Therefore, by creating a ranking list across all the cross-validation runs, features that are truly important and in the top 5 should appear frequently.

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

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

Table 1: Mean Imputation and ReliefF

Table 2: Class Mean Imputation and ReliefF

Table 3: Feature Ranking List