Lachlan Sinclair

3/9/2020

Assignment 2

Reflections

a. Accuracy by itself simply lets you know the number of true positives and true negatives as a percentage of the total number of instances. The accuracy however doesn't account for other possible issues that may cause its value to no longer be an accurate representation of the status of the data. When the class distribution is relatively equal the accuracy does reasonably well, however once the class distribution becomes skewed accuracy no longer provides meaningful insight. Accuracy also doesn't distinguish between false negatives and false positives, or true negatives and true positives. As the lectures and the textbook explain, these values should almost never be treated the same. One way to account for all these aspects are to use the expected value to compare predictors. Looking at the results for these assignments it is clear that the highest accuracy doesn't coincide with the highest EV value. Looking at the raw section, when the "balance in" field was reduced the accuracy went up while the EV suffered greatly. The EV provides far greater insight into how beneficial or costly certain predictors will be when compared to one another. However, the computation for EV requires additional work along with often hard to get knowledge (the b(PP,P) etc. values).

b. For both the mismatched training/testing section and the compensated training section, the PPV, NPV, markedness, F, MCC, and EV experienced similar large magnitude deltas when the balance of positive and negative instances were alerted for their respective sets (along with the compensation). The rest of the metrics stayed the same or experienced far less change in both sections. From a practical sense this means that both the mismatched train/testing section and the compensated training section predictors begin performing more poorly as the imbalance of positive and negative instances in the training or testing set was increased. Without going through the details of each of the altered metrics, overall as the balance of instances got worse in the testing and training set the metrics began to shift in the direction indicating that the predictors were performing poorly. The benefit of the compensation can be seen by comparing the raw section to the compensation section, the compensation allowed its performance to match that of an imbalance exclusively occurring in the testing set. For example, the EV value which provides valuable insight into the predictor's performance takes a significant nosedive in both sections buts stays almost even with each other. Overall, having an imbalance in the positive and negative instances in the training set when compensated for has roughly the same negative effect when the testing set is imbalanced.