Assignment 3 Report

Shala Chen 20698485

MED Classifier

I used sample mean of the training set to train the classifier, and there were 82 training samples erroneously classified by this classifier. I think it's because we only used sample mean and the boundary was a straight line so that there were many samples being misclassified.

Boosting

MED Classifiers

Before using boosting method on MED classifier, the error rate was 0.2050. To Apply boosting method, I selected 1/4 of training data as d1 and used d1 to train c1,then classify all training data using c1 and collected those misclassified data. I randomly selected half of the wrong samples and same number of correct samples to form d2, then use d2 to train c2. I tested all data in c1 and c2 and find those had different results as d3, and then trained c3 using d3. Finally I test all data sets in c1,c2,c3 and get the majority vote as the final results.

This time, the error rate varied around 0.1950.The boosting algorithm didn't significantly improve performance of the MED classifier.

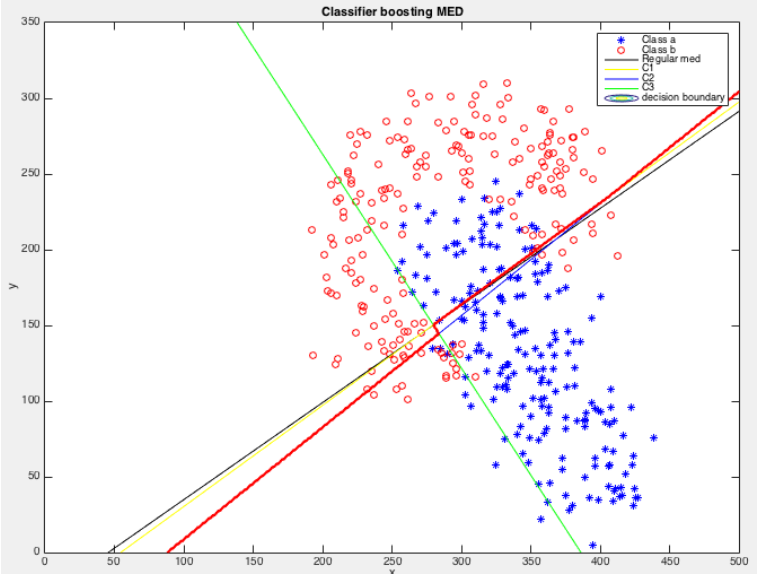

Figure 1: MED and boosting MED classifier boundaries

I think the reason why the boosting algorithm didn't significantly improve the performance of MED classifier is that c1 and c2 classifiers were so similar that their voting were almost the same, which hugely influenced final majority voting. That's why the final majority voting result was almost the same as c1 and c2, which also, made c3 meaningless for the most of the time.

Selected-Linear Classifier

When q=10, the error rate for regular selected-linear classifier is 0.2175. After applying boosting algorithm, the error rate decreased to 0.1525.

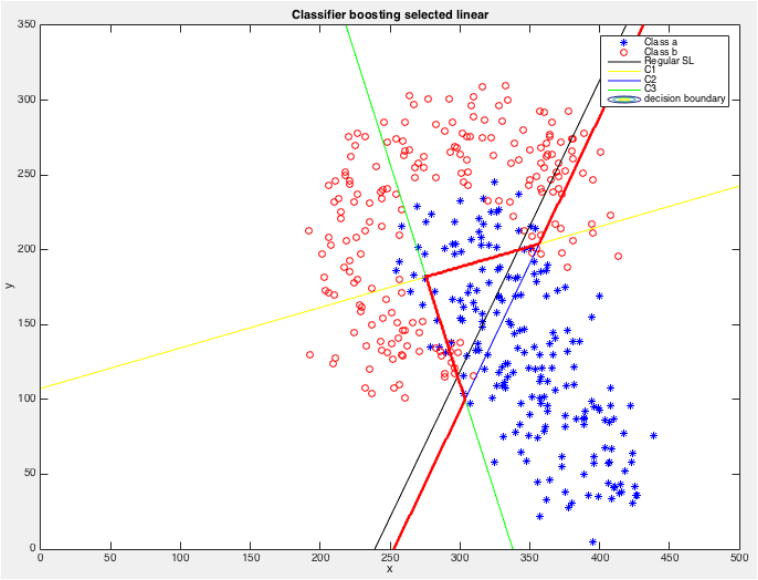

Figure 2: Selected-linear and selected-linear boosting classifier boundaries

Table 1: Error rates of classifiers

Classifier	selected-linear	c1	c2	c3	boosting selected linear
error rates	0.2175	0.2250	0.2200	0.3025	0.1525

From the plot we could find that c1, c2, c3 boundaries were completely different from each other, and their probability of classification error varied as well.

When I used selected-linear classifier, I randomly selected 10 data samples and picked the one with the lowest error rate as the training sample. Since I only used 1 sample to train the classifier, the boundary of c1, c2, c3 classifiers varied a lot. The variance leaded to a reliable majority vote, so that all three classifiers c1,c2,c3 had equal impacts on the final decision, which decreased the final probability of classification error.

Choice of q

For the sensitivity of the probability of error for the boosted selected-linear classifier method to q, I think it depends on the actual training size that we can use.

If the training set is very large, we can choose a large q so that we will have a chance to find a data sample with the lowest error rate. However, when the data set is very small, a large q will be useless because it will be choosing the same data sample over and over again and this could be a waste of time.

To get a reasonable result, q had better be smaller than the whole dataset, however, to be able to get a training set with a lower error rate, I think it's better to try at least half of the dataset,

which means q should be larger than half the size of the dataset.

Comparison of Probability of error

Table 2: Comparison of Probability of error

Classifiers	MED	boosting MED	Selected-Linear	boosting Selected-Linear
Probability of error	0.2050	0.1950	0.2175	0.1525

As I discussed above, due the similarity of the MED classifiers c1 and c2, the boosting algorithm didn't significantly improve the MED classifier. On the other hand, due to the variance of the Selected-linear classifiers c1, c2 and c3, majority vote worked great and the boosting algorithm significantly decreased the probability of error.