Battle of the SVMs

Theodore Wiebold

University of Texas at Dallas

The analysis started with the raw earthquake data that was transformed from a “.tsv” file to an excel spreadsheet. After initial testing, which will be discussed in a moment, the data was converted from 33.3 percent test, 66.6 training to 50 percent test and 50 percent training. The number of positive and negative events i.e. major earthquake vs. minor earthquake, had equal sample sizes in the second iteration of the processed data. In total, there were 47 major and 47 minor earthquakes in the training data and 46 major and 46 minor earthquakes in the testing data.

The initial testing of the raw data compared SVM with radial Bayes function and SMO solver to SVM with radial Bayes function and L1QP solver. There were also optimization trials for both models to see if their performance could improve. All models resulted in the same outcomes i.e. when the training data was used for training the misclassification rate was 18.01 percent and when the testing data was used for training the misclassification rate was 25.18 percent. The performance of the SVM models outperformed the Euclidean distance (28.78% error), dynamic time warping unconstrained (27.34% error), and dynamic time warping constrained (28.06% error), reaching a better error rate than the default error.

These results are misleading before turning to the confusion matrix of the prediction performance. The confusion matrix shows that there is heavy bias towards the negative outcome in the model when trained with the training data and a medium bias when trained with the testing data. Both models predict that all of the earthquakes in the data used for testing, regardless if it was the training or test data, were minor earthquakes resulting in the misclassification error rate i.e. in the training data 25.18 percent were major earthquakes and in the testing data 18.01 percent were major earthquakes. This means that the accuracy of the models were 50 percent since it got all the minor earthquake predictions correct, but all the major earthquake predictions wrong.

The models were both classified as soft margin SVMs due the size of the box constraint. The box constraint is the constant that controls the weight of the slack variable. Like Dr. Michael I. Jordan advises, the box constraint for both models were very close to 1/n e.g. 1/322. Since the box constraint was so small, misclassification was deemed unimportant. Hard margin was not considered feasible for this data.

Next, the processed data was analyzed in the same way as the raw data. Similar to the raw data, the models, the confusion matrix, and the ROC curve with area under the curve showed 50 percent accuracy. However, with the optimization trials for the models reduced the bias to almost zero and improved accuracy up to around 60 percent accuracy (ranging from 52 percent to 68 percent) depending on the optimization sample for the trail.

Comparing the scatter plots of the data revealed a difference between the two methods. The raw data grouped the data into four different groups as opposed to the processed data reducing the size of one of the four groups to almost nothing. This could have possibly been due to eliminating data points from the raw data set.

In conclusion, soft margin SVM did have the smallest error rate out of hard margin SVM, soft margin SVM, Euclidean distance, dynamic time warping unconstrained, and dynamic time warping with constraints, performing minimally at the default error rate.