**Ensemble Machine Learning Approach to the Practice Fusion Diabetes Classification Challenge**

Jeffrey DeVince

CPE-695 Applied Machine Learning

Final Project

5/13/15

**Introduction**:

The Practice Fusion Diabetes Classification Challenge [1] was completed using an ensemble machine learning approach. This challenge held in 2012 provided 10,000 de-identified electronic medical records and the goal was to build a model to predict the probability that a patient had type II diabetes. This challenge was created because diabetes is a pressing concern on the healthcare system. With earlier prediction of diabetes through analyzing electronic medical records, it is possible to fully test and possibly diagnose patients sooner.

**Methods**:

The dataset titled “trainingSet” provided at <https://www.kaggle.com/c/pf2012-diabetes/data> was used for both training and testing the machine learning model. Patient records 0-6500 were used as the training dataset and patient records 6501-9948 were used as the testing dataset. The data provided included a vast amount of raw information collected in patients’ electronic medical records such as allergy, diagnoses, medication, lab test, patient characteristics, and other data. Certain information was removed in order to make the challenge more difficult. Diagnoses indicating diabetes or diabetes complications, medications prescribed for diabetes, and lab tests for glucose or insulin were removed.

The feature selection process was one of the most important steps towards developing the model. Since so many different data types were provided, not all data types were used as features. The first features that were decided upon were based upon common risk factors for diabetes. This meant having the person’s age and person’s body weight index (BMI) included as features. The age was given directly through patient’s year of birth, so that was used directly as the feature. For the BMI, a BMI measurement was recorded each time the patient met with a doctor, so there were multiple measurements for BMI for a single patient. In order to have a single BMI feature, the median of all non-zero BMI measurements was used. Similarly to the BMI data, the height, weight, systolic blood pressure, diastolic blood pressure, respiratory rate, heart rate, and temperature were given in the same format. Therefore, the median of all non-zero values for each of these data types was used as a feature. The last three features that were used were based on medication that the patient was taking, diagnoses of the patient, and lab tests that the patient received. Data for these three items were available in similar formats and implemented using the same method. The format of these data was that every instance, such as every time a lab test was order for that patient, was available with a link to the specific patient. This means that if a patient had a certain lab test once every six months, it may appear six times in the data. In order to account for the amount each item appeared, the feature used was the count of how many times a specific medication, diagnosis, or lab test occurred in the data. Since there were many different types of medications, diagnoses, and lab tests in the data and each would require its own feature, it was desired to remove any items that would not be useful in training the random forest model in order to improve computation time. However, it was important that no useful data would be removed. Therefore, for each data type, specific medications, diagnoses, and lab tests were only removed if there were less than ten mentions in the 10,000 records. This process removed rare items that did not have enough instances to help prediction. By applying this removal process, the number of features was decreased from 2553 to 1478 for medication, 3902 to 2304 for diagnoses, and 340 to 219 for lab tests.

Once these features were decided, the machine learning approach was needed. It was decided to use an ensemble method in order to hopefully improve the prediction ability over the prediction ability of a single machine learning algorithm. It was decided to combine a random forest, support vector machine, and Gaussian naïve Bayes algorithm. These were decided because they are three of the best and most widely used machine learning algorithm. By combining these three algorithms, it was hoped that the disadvantages of any one algorithm would be minimized. The method of combining the three algorithms was to train each fully on the training set, and then each predict the probability that a patient has type II diabetes for each patient from the dataset, and average these probability values for each patient.

The prediction from the model for each patient’s data was compared to the ground truth using a log loss formula to quantify the performance of the model. The log loss was the primary quantification used to evaluate the models in this competition. Log loss is defined as:

where N is the number of patients, log is the natural logarithm, (y\_i ) ̂ is the posterior probability that the ith patient has diabetes, and yi is the ground truth where yi = 1 indicates diabetes and yi = 0 indicates no diabetes.

**Results**:

The testing was performed for every combination of the three algorithms and evaluated based on the log loss as was done in the actual competition. The combinations performed were random forest alone, SVM alone, naïve bayes alone, random forest and SVM, random forest and naïve bayes, SVM and naïve bayes, and all three. The comparison between these different combinations is shown in Figure 1. The naïve bayes alone is not shown because it had a log loss value of 9.315.

Figure 1. Comparison between the different algorithms.

It was found that the random forest alone had the best log loss value with a value of 0.38294. This would have placed the model at 77th place out of 146 teams in the actual competition as shown In Figure 2.

Figure 1. The log loss values for the best submission from all teams in the competition. The bottom four teams were not included because they had log loss values above 1.

**Discussion and Conclusion**:

Overall, it was found that the random forest algorithm by itself achieved the best performance. This was an unexpected result because it was expected that the combination performance would be better than any individual algorithm. Both the random forest and the SVM algorithm performed the task well, but the naïve bayes did not work well for this dataset. The random forest model would have been very competitive had it actually entered the competition. The 77th place result has a caveat because the model in this paper only trained on 6500 records while the models in the competition were trained on the full 9,948 records since there was a separate testing set that was unavailable for this paper. This means that the model in this paper achieved 77th place with significantly less training data. It is assumed that this position would be improved if the entire 10,000 training records were used.

**References**:

[1] Practice Fusion Diabetes Classification. July 2012 – September 2012. Hosted by Kaggle Inc. <https://www.kaggle.com/c/pf2012-diabetes>