Predicting Parkinson's Disease Severity from Patient Voice Features

Nicolas Genain*, Madeline Huberth†, Roshan Vidyashankar†
*Department of Management Science and Engineering, Stanford University
†Center for Computer Research in Music and Acoustics, Stanford University

***Abstract*—This paper describes the machine learning methods and modeling used to predict continuous measures of Parkinson's Disease Severity from voice recordings of patients. Two datasets are analyzed. Bagged decision trees (random forests) resulted in an improvement on the previous model accuracy for one dataset, predicting severity measures at 2% accuracy on a 0-176 scale. Other methods are described for both datasets, as well as the limitations of each.**

I. Introduction

Parkinsons disease is a progressive disorder of the central nervous system that causes loss of control of movement. Early symptoms include tremors in hands, and slowed, slurred speech, with symptoms developing into more uncontrollable, full-body tremors. Changes in memory and cognition also occur. As many as 6.3 million people live with Parkinsons worldwide [1], making it the second most common neurologic condition after Alzheimer's disease. While many diagnosed with Parkinson's can go on to live for many more years, their quality of life is diminished.

There is no cure for Parkinson's, but it may be possible to slow down or even prevent the progress of this disease by detecting it early. There are various methods of brain imaging for early Parkinsons detection, but they are expensive [2]; there is a need for inexpensive, scalable diagnostic techniques. Recent developments include tracking postural gait and sway [3] and identifying anomalies in speech recordings of patients [4], [5]. This second technique may prove sufficiently scalable. Dr. Max Little and PatientsLikeMe, an online communication platform for PD patients and simultaneous research platform, have joined with Sage Bionetworks to collect data using internet-based voice recordings, with the intention of proving that crowdsourcing approaches can potentially predict self-reported PD severity measures. Their dataset was released to us as part of a dry-run for their public competition which is to be launched this coming summer.

This paper describes the models and methods we used to predict PD severity measures from voice features on the Synapse.org dataset, and also on the University of California Irvine 'Parkinsons Telemonitoring Dataset', another dataset containing PD severity measures and voice features, though non-internet based. Reasons for analyzing both datasets are described, as is a reflection on how our work contributes to the scope of the larger Synapse.org project.

II. University of California Irvine Parkinson's Telemonitoring Dataset

The UCI Parkinson's Telemonitoring Dataset is the result of a six month trial on 42 Parkinson's Disease patients. Recordings of sustained vowel phonations were recorded weekly by the patients over the course of 3-6 months from Intel Corporations telemonitoring system, the At-Home Testing Device (AHTD), which is designed to track PD progression including speech tests. This dataset gives us a high confidence in the quality of the voice features. Furthermore, it includes three clinician-evaluated records of the patients' full UPDRS scores (Unified Parkinson's Disease Rating Scale) This UPDRS score was recorded by a physician on three occassions: at the beginning of the trial, after 3 months, and after 6 months. Then, at each timestamped voice recording, the linear interpolation of the pdrs between the real values was associated.

In total, there are approximately 6,000 recordings from the 42 patients, a sizeable dataset compared to other datasets of voice features from PD patients. There are 16 voice features - acoustically extracted measures of the patient's recordings - available in this dataset. Direct recordings from the patients are not publically available.

The research group responsible for organizing this dataset reported model accuracy within approximately 7.5 points of the total UPDRS score, which in this dataset can be evaluated up to 176 points, though the highest reported value was 55.

A. Random Forests, and Moving Average Inclusion

R's randomForest package, which implements Leo Breimans Random Forests [6], were used to predict the interpolated UPDRS score. In Random Forests, a group of decision tree classification or regression trees is trained and bagging then combines the predictions. Each tree selects a portion of the features at random as well as a random resampling of the data to train on. The three parameters that were modified were the number of trees trained, and the number of variables randomly sampled at each split of the tree.

Inclusion of all of the features to predict total-UPDRS score resulted in a best model with an RMSE of 8.38. The following parameters were used to obtain this tree: 506 trees, 8 features included at each split of the tree.

Both ridge regression models and LASSO models were fit as feature selection methods, using the glmnet package in R. We split the samples into a training set and a test set in order to estimate the test error of ridge regression and the LASSO. For both models, ten-fold cross-validation we used select the best value of the tuning parameter λ. The lowest test RMSE associated with ridge regression and the chosen λ was 10.627, and for LASSO was 10.31, which are both an improvement on simply fitting a model with just the intercept, for which the test MSE is 10.78. However, since the LASSO resulted in such a substantial improvement in MSE, even over the ridge, the twelve features selected by this model were included in subsequent random forests. As a direct comparison, running a Random Forests model with only these features results in an RMSE of 8.33, a small improvement on when all features were included.

Moving Average. One advantage of this dataset is that it includes time-series data for each patient. To see if slightly smoothed voice measures improved our Random Forests model, we calculated a moving average variable for each of the LASSO-selected features (testing a range of moving average windows from 2-20 samples), and included those new features with the original LASSO-selected voice features in our Random Forests testing.

This tactic resulted in us beating the accuracy of the work of the original researchers. We tried moving averages centered around the predicted UPDRS point, and also solely a lagged moving average. While the centered window performed only slightly better than the lagged window, the results are very comparable. The best performance this tactic was achieved using the parameters of 502 trees, 8 features included at each split of the tree, used a centered window, and resulted in an RMSE of 1.968, which is a great improvement on our prediction in comparison to not including the moving average features.

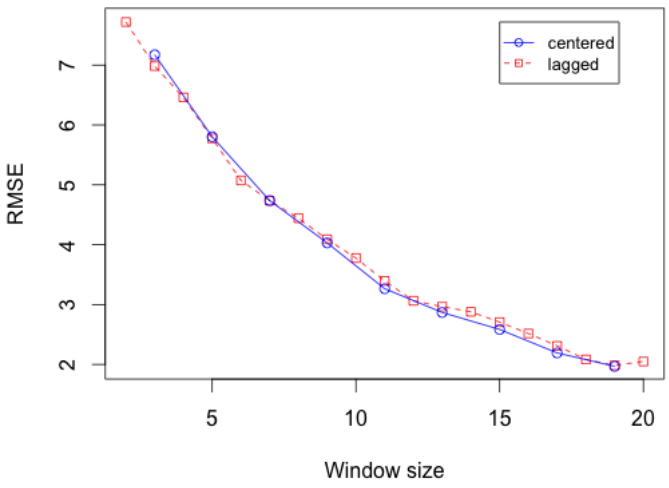

Fig. 1. Comparison of improvements in Random Forest RMSE based on window size of moving averages. The window for the moving average was both centered and solely lagged.

Also included in the randomForest package is information about the importance of the predictor variables, which includes a measure of how much MSE increases when that variable is randomly permuted while all others are left unchanged. For the case of our most successful centered-window moving average random forest, the most important 12 predictors were all the averaged features. This confirms that smoothing the features considerably improves the model, over including only the regular, more noisy, non-smoothed features in the model.

B. Linear Regression

All the subsequent methods have been implemented with the scikit-learn package for python. Substantially, they only made very few improvements on the prediction and were a lot less effective than the Random Forests. One of the first things we tried is to predict the PDRS score with a linear regression. Two types of regression have been tried: a regression per patient and a regression on the whole dataset. In all cases, the data has been preprocessed to have mean 0 and variance 1. We have also used LASSO, Ridge and Linear selection to test the importance of each feature and select the most important ones. Given the size of the dataset, we were able to implement a Leave One Out Cross validation too. As a baseline, we first took the prediction by the mean PDRS score 29.02 over all the data. This gave a RMSE of 10.70. A regression on the whole dataset yielded worst

results: a RMSE of 10.90. For the personalized linear regression, quick plots have shown us that this method wasn't consitent. For some patients it worked admirably, for others it literally predicted the opposite of the trend. The following two graphs will give the intuition why that happened. The green dots are the PDRS score and the blue dots one of the features. The red line is the linear regression of that feature. We can see that in the first graph, the linear trend of the feature goes is strongly correlated with the pdrs trend, but, in the second graph, this same features is negatively correlated to the trend. In the third case, the features are totally unpredictive of the pdrs score.

C. Support Vector Regression

We used this technique. We did a grid search on the parameters C, Epsilon, gamma and the type of Kernel. At best, we had a RMSE of 10.2 which is a very scarce improvement on the prediction by the prediction by the mean. As a result of our initial poor results, we also implemented a classification problem with SVMs and a grid search too. This gave better results: we were able to predict the right PDRS score to the right bucket 50% of the time. Nevertheless, by looking at the errors, we saw that whereas most of them were off by only one bucket which isn't a huge problem, a significant number of them were predicted further than 2 buckets. This hampered our confidence in the results.

D. Hidden Markov Models

By trying further to simplify the problem, we also tried to predict the trend of the PDRS between two measures given the features. This gave two possible states : "getting better" and "getting worse". We would first train the model on the data seen and then, given the estimation, predict the sequence of states between "getting better" and "getting worse". As in the linear regression, we trained one model on all the patients and also one model for each patient. We also found similar results to the Linear Regression: some patients' predictions were very accurate but others were totally off.

III. Patient Voice Analysis Dataset

Originally, we started working with a different dataset, one not yet released to the public. This section includes description of this dataset, our analysis and work with it, and a discussion of its limitations and why we began working with the UCI dataset.

The Patient Voice Analysis (PVA) dataset contains voice recordings of voice phonations (3-30 seconds

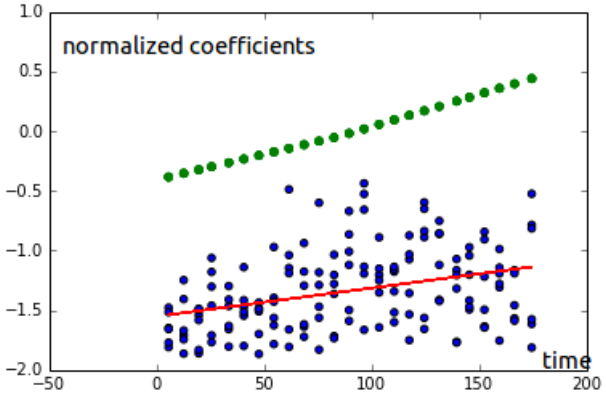

Fig. 2. Linear interpolation of the features vs the pdrs score in a good case

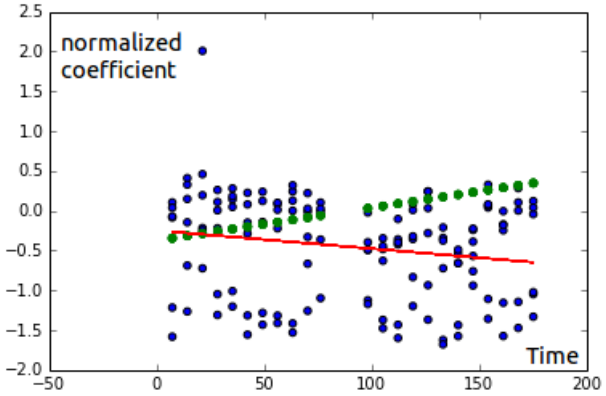

Fig. 3. Linear interpolation of the features vs the pdrs score in a bad case

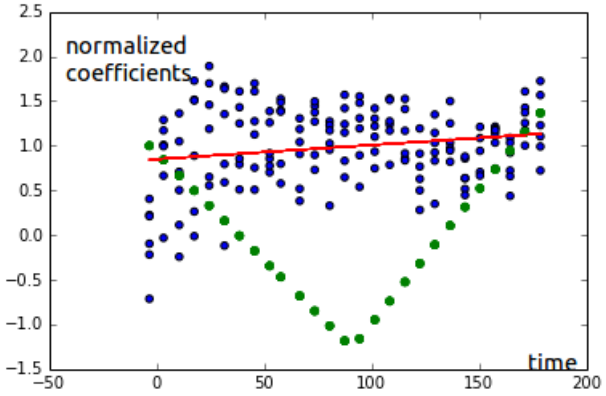

Fig. 4. Linear interpolation of the features vs the pdrs score in anoter bad case

long of a sustained voiced ‘ah’), self-reported symptom assessment (PDRS - Parkinson’s Disease Rating Scale as well as Hoehn & Yahr stage classification) and demographic information about the caller.

The PDRS scale, in this dataset, is a partial version of the clinically accepted, widely-used UPDRS score, in which the answers to 17 questions about the patient’s symptoms (answerable on a 0-4 scale) are scaled and summed to be on a 0-100 scale.

The training data set also includes 38 features extracted from the voice recordings. 26 of these features were cepstral coefficients and their derivatives, while 5 features dealt with standard audio features (fundamental frequency and RMS power measures), while three features were developed previously to characterize voices of Parkinson’s patients [7], [8].

Each row in the dataset corresponded to one report from a Parkinson’s patient. There were 365 users total, with some repeat calls, for a total of 390 rows. There was little to change to be able to work with the dataset. There were 15 rows which had no voice feature data available (the call quality was especially poor), and our team removed them from all analyses.

While there is more data available, it was held out by the providers of the dataset as a test set, eventually to be used to test competitor-submitted models in the public competition.

Our goal with this dataset was to predict the PDRS score from the voice features. Below we present various strategies.

A. Distribution of dataset features

The mean PDRS = 21.28, and variance = 122.57. The variance is the baseline predictor we choose: we use as measure the Root Mean Square Error with the PDRS, which is 11.07. Predicting the mean is the most nave way of predicting the PDRS-score. We will try to improve upon 11.07. See Figure 1 for the distribution of PDRS scores.

B. Support Vector Regression

As with the first dataset, we implemented a grid search on the SVR with a selection of features to find the best prediction. In this case, we arrived at a variance of 108 which is a slight improvement on the baseline prediction.

C. Prediction of the Hoehn and Yahr scalse

We also implement a SVM for the Hoehn and Yahr scale. We had pretty good results : error rate of 38%. And the when there were errors, they weren’t too far away from the reality.

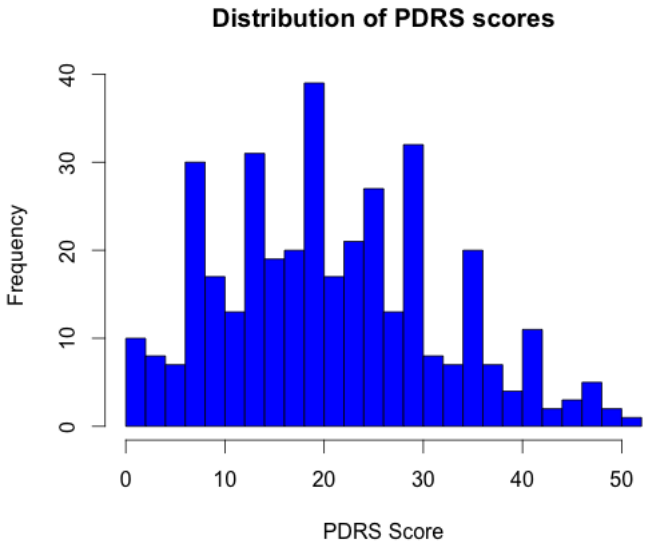

Fig. 5. The distribution of the PDRS scores is fairly Gaussian.

D. Random Forests

Similarly as for the other dataset, three parameters were modified for the Random Forests calculation (again, using the randomForest package in R): the number of trees trained (a range of 500-3000 was tested), the number of samples to include in the training of each tree (a range of 300-350 samples was tested), and the number of variables randomly sampled at each split of the tree (a range of 10-15 was tried). The value of RMSE incuded in the randomForest package was used. The smallest RMSE from these tests was 10.71, with the following parameters: 503 trees, 12 number of variables at each split, with 322 samples to build each tree.

E. K-Means Clustering

We wanted to try an unsupervised learning method on the data. Instead of thinking of the problem as a regression problem, we tried to cluster the the data into buckets with similar PDRS scores. The K-means clustering algorithm was used for this. We tried several iterations, varying a number of parameters each time - number of clusters (4 - 12), method of initialization (random, K-means++) and sets of features (all features, audio features only, PCA-selected features). To evaluate our results, we used several standard metrics - Homogeneity and Completeness Scores, Silhouette coefficient [10] and the Adjusted Rand Index [11]. Clustering was not very effective. We had low correlation scores for all of our runs. Using the k -means++ initialization method, 4 clusters and all audio features, we got a low silhouette score of 0.119 indicating poorly defined clusters.

F. Data Limitations

As we worked with the dataset, some of its limitations were revealed, both uncovered by us and by the Synapse.org PVA team.

There was an issue with the self-reported PDRS scores. While some of the repeat callers have consistent scores, some have a potentially problematically large range. For instance, one participant called three days in a row, and during these days, reported scores of 25, 27, and 19. Also noted is that two 0's in repeated PDRS score come from a user who on his first call, reported a PDRS score of 23. The possibility of unreliable data, or data which may not be consistent with the actual condition of the disease bears weight on the overall possible reliability of our model.

We were also notified by the Synapse.org PVA team that they felt the extracted voice features by several methodological issues. The audio was collected through the Twilio voice API, and so were not of high quality, limiting their use as predictors.

Furthermore, the dataset was small. Given the speed with which data like this can be collected - our data was collected over the course of only two weeks in January 2014 - more data could have improved the predictive power of our models.

We conjectured that including some healthy controls would improve our model, so we recruited volunteers who perform the 10-second vowel phonation, which we recorded with a Zoom H2 Handy recorder. After elimination of those volunteers who had unusable data (mostly due to wind noise), we included 20 healthy controls in the dataset, calculated the MFCC's and recalculated the MFCC's for the original data, and tested if an SVM classifier could predict which were the controls. It had perfect accuracy, more than likely due to the difference in audio quality. While we could have reduced the audio quality of our recordings to be comparable to those made through the Twilio API, given the other limitations of the dataset, we chose to move our analysis to a more optimal dataset, suggested by a member of the Synapse.org PVA team.

The last limitation of the dataset that we found is that the features are highly correlated. We ran an SVD on the features matrix and found that 99% of the energy of spectrum was contained in the first eigenvalue. Almost all features were linearly related. We plotted 6 here the first vs the second feature.

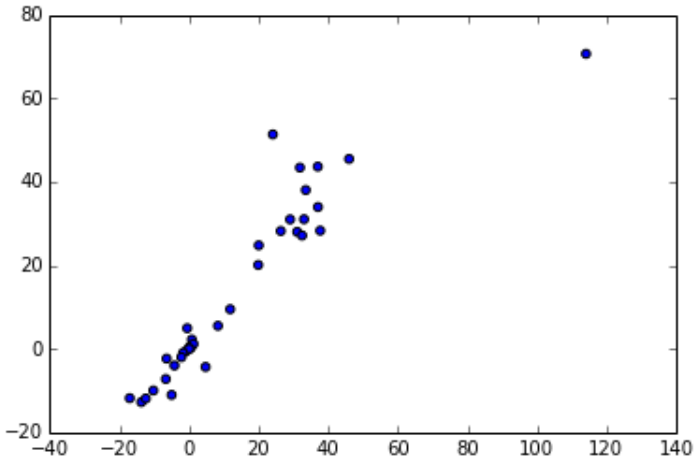

Fig. 6. Plot of the first vs the secon feature.

IV. Discussion

Though both datasets had their limitations, we were pleased with our results for the UCI dataset - the original prediction from the researchers who put together this dataset was an RMSE of 7.5. By including moving averages to a Random Forest model, we improved model accuracy to an RMSE of 2. While this strategy resulted by far in the biggest gains in accuracy, smaller gains were made by tweaking the parameters of the Random Forest model, as well as using feature selection methods, specifically, LASSO.

Certainly our results would be even more convincing if the UPDRS score associated with every voice recording was validated by a clinician. Due to likely numerous reasons (e.g. expenses, convenience), only 3 UPDRS scores per person were evaluated by a clinician, taken at the beginning, middle, and end of the treatment. The rest of the scores were researcher-created linear interpolations between these three scores. In reality, the UPDRS scores are likely much more noisy; and working with real measures would have resulted in a much more meaningful model.

This limitation from the UCI dataset aside, we are far more confident in our models of the UCI dataset than we were with the original dataset of this course, the PVA dataset. While our models resulted in a prediction accuracy that slightly improved the baseline prediction, these gains were slight. Over the course of working with the dataset, it became clear to us that the quality of the data itself likely limited the potential prediction accuracy of the model - PDRS scores were user-reported and errors in reporting existed, and the quality of the voice recordings from which the features were drawn was poor. Furthermore, the dataset was small.

However, we discussed our perception of these limitations (as well as our gains in prediction accuracy with our models) with the PVA Synapse.org team, and are hopeful that our insights and commentary prove useful

to them as they make final changes before opening this dataset to a public competition.

V. Future Work

For furture work, we need to do the following:

- Try random forests with moving averages on the PVA dataset in order to see if it can help get better results on this data.
- Check for over fitting - we only have 42 patients. so we can check on more patients.
- Verify on non ill patients that the PDRS score is zero (or very low).
- There is also a great work to do in collecting more real PDRS data points. Right now we use interpolated PDRS scores. We should have the real PDRS score for each voice recording in order to be confident in our prediction capabilities.

VI. Conclusion

This paper presented our work on predicting the accuracy of Parkinson's Disease severity from voice features. Our most successful accuracy resulted from including voice features smoothed by a moving average in a Random Forest model using the UCI dataset.

The body of research that predicts PD severity from voice features is moving in a direction such that, perhaps in the near future, PD patients will be able to monitor the progression of their PD by simply recordings their voice, without the need for a trip to the clinician. While several hurdles remain to implement this kind of technlogy at scale (as evidenced by the PVA dataset), the outlook seems promising.

We are pleased to have contributed to the PVA project, and to have improved the prediction accuracy on the UCI dataset.

References

[1] "About Parkinson's Disease." Facts about Parkinson's Disease Neurological Movement Disorder. Web. 17 May 2014.

[2] University of Haifa. "Early detection of Parkinson's disease by voice analysis." ScienceDaily. ScienceDaily, 24 May 2010. www.sciencedaily.com/releases/2010/04/100419102927.htm.

[3] Arora, S., Venkataraman, V., Donohue, S., Biglan, K.M., Dorsey, E.R., and Little, M.A. (2014) High accuracy discrimination of Parkinsons disease participants from healthy controls using smartphones in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing, 2014. ICASSP 2014 Proceedings*: Florence, Italy.

[4] Tsanas, A., Little, M.A., McSharry, P.E., and Ramig, L.O. (2010). Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinsons disease symptom severity. *Journal of the Royal Society Interface*, 8(59):842-855.

[5] Tsanas, A., Little, M.A., McSharry, P.E., and Ramig, L.O. (2010), Enhanced classical dysphonia measures and sparse regression for telemonitoring of Parkinson's disease progression in /emph2010 IEEE International Conference on Acoustics, Speech and Signal Processing, 2010. ICASSP 2010 Proceedings. Dallas, Texas, USA. pp. 594-597.

[6] L. Breiman (2001). Random forests. *Machine Learning*, 45(1):5-32.

[7] Little, M.A., McSharry, P.E., Hunter, E.J., Spielman, J., and Ramig, L.O. (2009), Suitability of Dysphonia Measurements for Telemonitoring of Parkinson's Disease in /emphIEEE Transactions on Biomedical Engineering, 56(4):1015-1022.

[8] Tsanas, A. Little, M.A., McSharry, P.E., Hunter, E.J., Spielman, J., and Ramig, L.O. (2012), Novel speech signal processing algorithms for high-accuracy classification of Parkinson's Disease *IEEE Transactions on Biomedical Engineering*, 59(5):1264-1271.

[9] Tsanas, A. Little, M.A., McSharry, and Ramig, L.O. (2010). Accurate Telemonitoring of Parkinsons Disease Progression by Noninvasive Speech Tests in *IEEE Transactions on Biomedical Engineering*, 47(4):884-893.

[10] Peter J. Rousseeuw (1987). Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. Computational and Applied Mathematics 20: 5365. doi:10.1016/0377-0427(87)90125-7.

[11] Hubert, Lawrence and Arabie, Phipps (1985), Comparing partitions in Journal of Classification pp. 193-218 doi:10.1007/BF01908075