Parkinson Disease Classifier Using Patient Voice Recording Data

Han Lee, Yonghyun Ro, Changwhan Yea

Department of Computer Science, Stanford University, Stanford, California

[hanlee, yro, yeaz]@stanford.edu

Abstract

Patients of Parkinson's disease (PD) regularly need to visit doctors for time-consuming examinations to track their progressions. This process can be simplified if we can predict their severities through voice recordings. With 389 Parkinson Patients' voice recordings, we developed a classifier that can predict each patient's Hoehn Yahr score and PDRS score, which are two main indicators of severity. We achieved 62.2% accuracy in predicting Hoehn Yahr score and RMS error of 9.57 in predicting PDRS score. Voice filtering algorithms such as cepstral mean normalization, band path filter, and spectral subtraction were applied to pre-process the data and then combinations of machine learning algorithms such as support vector machine and decision tree classified each recording to the severities.

Index Terms: Parkinson's disease, disease severity prediction

1. Introduction

Parkinson's disease (PD), also known as the hypokinetic rigid syndrome, is a degenerative disorder of the central nervous system. According to the Parkinson's Disease Foundation, 10 million patients around the world are diagnosed. While 60,000 people are reported to be diagnosed in America every year, an estimated 20% of people with Parkinson's are never diagnosed. This reflects the difficulty behind identification and medical treatment of the disease, which currently has no cure and no blood test to detect it for early intervention. Tracking PD symptom progression currently requires the patient to physically visit a clinic and go through time-consuming examinations by trained medical staff.

While current methods of symptom monitoring are inconvenient and costly for large-scale implementation, research has been conducted to rapidly screen PD with voice recordings of patients. Voice impairment has been identified as one of the earliest indicators and common symptoms of PD, as studies report 70-90% prevalence after the onset of the disease and 29% of patients considering it to be their greatest hindrance. Max Little, the head of Parkinson's Disease Initiative, has introduced a method that can provide a clinically useful estimation of whether or not a person has PD. This method listens for three main clusters of symptoms in the voice - vocal fold tremors, breathiness and weakness - as well as the way the jaw, tongue and lips fluctuate during speech, and is reported to have 99% accuracy rate on a dataset of 42 patients and 5,923 sustained phonations.

Based on these previous studies, we aimed to create a classifier that extracts features from voice recordings of Parkinson's disease patients and detect the severity level of the disease. Given simple voice recording data of patients over the phone, our goal was to clean up the recording with appropriate preprocessing methods, extract vocal features from the data, and run various data mining models with the feature set to predict the final severity score.

2. Data

We used the dataset provided by Synapse.org, which was consisted of 389 phone recording of PD patients, their Parkinson's Disease Rating Scale (PDRS) score, Hoehn and Yahr scale score, and demographic information, such as age, gender, years passed since PD diagnosis.

Both PDRS and Hoehn and Yahr scores are commonly used systems of measuring PD severity and determining where a patient might be in with regard to progression. A PDRS score ranges between 0-100, with 100 being most severe. It is composed of an evaluation of behavior, mood and a self-evaluation of activities of daily life, including speech, walking, and handwriting etc. A Hoehn and Yahr score evaluates severity based on five different stages, with 5 being most severe. Both scales are incorporated in the Unified Parkinson's Disease Rating Scale (UPDRS) scale, with assesses limitation of daily activities and non-motor symptoms in more detail. In our study, we used both PDRS and Hoehn and Yahr scores for classification.

2.1. Voice Data Filtering

All 389 raw data were the phone recordings of 8kHz sample rate, which included relatively high signal to noise ratio. We have applied three filters: cepstral mean normalization, band filter, and spectral subtraction. All possible combinations of below filters were used to find optimal accuracy.

2.1.1. Cepstral Mean Normalization

A common problem with speech recognition systems is that the characteristics of the channel may vary from one session to the next. Cepstral mean normalization minimizes the effect of such differences by subtracting the cepstral mean from each frame.

2.1.2. Band Filter

Human voice has frequency range from approximately 300 Hz to 3,400 Hz. Anything beyond the range could be regarded as a noise. We made the low-pass filter that filters out the noises that have frequencies over 3,400 Hz. However, as shown in the below diagram, intensity of signals in higher frequencies ranges are relatively small.

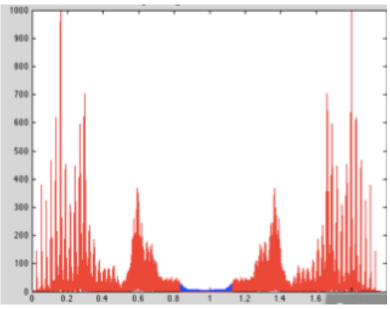

Figure 1: *Band filter plot in different frequency range*

2.1.3. Spectral Subtraction

One common way to reduce the background noise, prevalent in phone recordings, is to apply spectral subtraction. Among different spectral subtraction methods, we adopted power spectral subtraction as it is most suitable for stationary noises. Power spectral subtraction observes the initial 0.25 seconds of each recording before the patient starts speaking to estimate the power spectrum of noise. Such spectrum is subtracted from each frame to find the clean data.

3. Experiments

3.1. Setup

Because we were not sure if data cleanup and data cutting would help, we had to run three separate trials with different data sets for each experiment to determine if data cleanup improves our results. Below we describe the setup for each experiment, which was the same for all three trials.

For each experiment, data was randomized and 90% of data was chosen as a training set. In classification models, data was separated for each possible label, and 90% of each data set was sampled afterwards to keep the distribution the same.

Because the size of data set is small, we utilized 10-fold cross validation to determine the accuracy of the model. For feature selection and parameter tuning steps, results from five iterations of 10-fold cross validation with different training were averaged to minimize the randomness. Each experiment was designed to predict either the PDRS score or the Hoehn Yahr score.

For the PDRS score, each experiment was designed to output RMS score, and the results from previous researches were used for comparison. Since our data contains much noise as mentioned above, comparison step was used to decide whether or not that experiment is worth pursuing further.

As for the Hoehn Yahr score, since we are doing five-label classification, we used the ratio of correctly predicted labels to total labels to evaluate the correctness of each experiment. Also, we implemented random guessing, weighted random guessing, and naïve SVM classification as our baseline methods.

3.2. Early Approaches

3.2.1. Linear Regression

Table 1. *RMS Error by predicting all data points as average of train data*

Method	Error from other research	Error from our data
Chance	7.5	11.1
Lasso	6.9	N/A
SVR	5.9	N/A

We used linear regression as our earliest approach. Our baseline measurements are from the previous research that used clean data (Bayestehtashk and Asgari, 2013). Although the dataset is different, we believe that the magnitude of the improvement from Chance to Lasso should also be visible in our attempts if our cleanup attempts are successful.

Because the range of values is different for each feature, doing min-max scaling resulted in more accurate results. For feature selection methods, we attempted three different methods. The first was selecting features based on their weights computed from naive linear regression. The second method was selecting features as we did in the first method but select only those uncorrelated with features already selected. The last method attempted was forward-greedy feature selection algorithm. Also, we ran our experiments on four different types of dataset: The first is original feature set given by Synapse. The second is an expanded feature set of 1582 features that we extracted from raw recordings. The third is a set of the same 1582 features but with cepstral normalization, extracted from recordings that went through spectral subtraction. The last dataset we used was a set of 1582 features with cepstral normalization that we extracted from only the final five seconds of each recording. Our intuition was that tremors in speech would get worse when patients start running out of breath.

Table 2. *Best RMS for each data set with min-max scaling and greedy feature selection*

Dataset	RMS Error
Original	10.44
Expanded Feature Set with 1582 Features	10.35
Cepstral Normalized Feature Set with Spectral Subtraction	10.35
Cepstral Normalized Feature Set with Spectral Subtraction - only end considered	9.57

The magnitude of improvement in the last dataset is comparable to that of our baseline research result. While this is an interesting results, the number of data points we could consider given the criteria is few, only around 220 points, because some recordings were cut in the middle. We will need a bigger data set to confirm our results further in the future.

After our initial experiments, we attempted combining SVM with Linear Regression. We created two systems: the first system at first predicts whether the PDRS score of the given sample is over or under the mean PDRS score of the training set by using SVM, then will predict the magnitude of the PDRS score away from the mean for the given sample. The second system at first predicts whether the PDRS score of the given sample is around the mean PDRS score of the training set, more specifically, +-4 from the mean by using SVM, then will use a general linear regression model that predicts the PDRS score for those classified to be outside the range, and will use another linear regression model to predict the magnitude away from the mean for those classified to be around the mean. SVMs used Gaussian kernels, and went through parameter tuning to ensure the highest generalization accuracy.

Table 3. *Results from Two Systems*

Method	SVM Accuracy	Regression Error	Method error
SVM to determine sign	64%	6.5	12.1
LR to predict magnitude	72%	10.1 / 3.2	11.0

Since the SVM's accuracy was low for both systems, less regression error did not help in improving the final RMS error.

3.2.2. Decision Tree

Extending the number of features to 1582, we believed that some of feature would be a strong indicator for certain levels of severity. We, thus, manually went through the first 500 features to find the correlation between certain level of severity to the feature. For example, if one of the shimmer features of a patient is higher 0.85, then the patient is categorized with Hoehn Yahr score of 5. After iterating the first 500 features, we were able to identify the Hoehn Yahr score of 150 patients with 95% accuracy. Promising it seems, the accuracy was based on the training data to build the decision tree. Using the same decision tree to predict the remaining 224 patients in the test set, we achieved the accuracy of 36%. Because SVM showed a better accuracy, we decided to stick to classification by SVM.

3.2.3. Predicting Hoehn Yahr Rating

In predicting the Hoehn Yahr rating for a given patient, we deemed SVM to be the best approach because of its predictive power in binary classification. To obtain baseline measurements to compare results from our future endeavors against, we implemented three different methods. The first method is pure random guessing, from 1 to 5, and the second method is weighted random guessing based on label weights in the training data, and the last method is naive 1 vs 2 vs 3 vs 4 vs 5 SVM, using all features and using either polynomial and Gaussian kernels.

Table 4. *Baseline Accuracies for Predicting Hoehn Yahr Rating*

Method	Accuracy
Random Guessing	20%
Weighted Random Guessing	30%
Naïve SVM	38-42%

3.2.4. Support Vector Machine

Each data point has a discrete rating from 1 to 5, so we theorized that treating the problem as a classification problem would be the best approach. As our first attempt, we relaxed the problem from a quinary classification problem to a binary classification problem. If any two-class division can be done on the data with minimal error, the problem becomes much simpler as further classification can be done for each of the separated groups. To find the best two-class division from the initial state, we grouped data in different combinations of labels to see which clustering could be predicted with high accuracy. As seen by the graph that follows, 1,2,3 vs 4,5 division achieved the lowest error rate, less than 5% with the right selection of kernels and regularization parameter.

Assuming that we can predict which group of labels the data points belong to, the next step is to predict which rating they actually have to be given from their corresponding group of labels. At each step, we will be considering smaller and smaller groups. The following diagram characterizes this system.

There were many challenges in this approach. First, we had eight different SVMs in total for all classification stages, and every SVM had to be tuned separately in order to achieve the best results. Also, since we are not sure which kernels would work the best for each SVM, we had to run every SVM using three different types of kernels: quadratic kernels, cubic kernels, and Gaussian kernels. Lastly, since we were not sure which data filtering method would have positive effects, we had to compute optimal SVMs for each dataset for comparison.

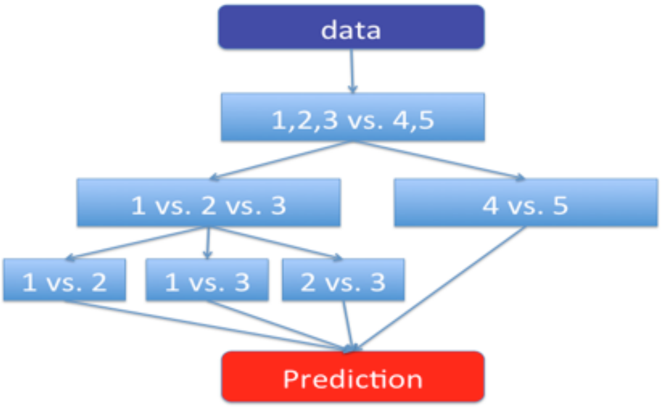

Figure 2: *SVM in Decision Tree Fashion for Hoehn Yahr Prediction*

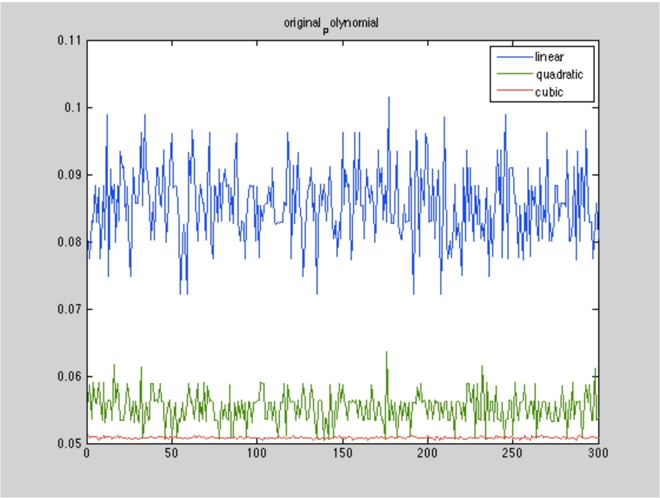

Figure 3: *Regularization parameter vs error rate for 123 vs 45 SVM*

Main methods of optimization were greedy-forward feature selection algorithm first and parameter tuning afterwards. Initially, we tried tuning just the parameters as SVMs, in theory, are resistant to over-fitting by choosing the right parameters, but we found that choosing the so-called right parameters is a difficult and time-consuming task; results computed by grid search were still below 65% for most of the single-label vs. pair-of-labels classifiers (However, it should be noted that this is not worse than random guessing because the data distribution is not 1:1:1). In contrast, using greedy-forward feature selection algorithm returned a classifier that is more accurate by more than 4%. Tuning the parameter afterwards, we were able to increase the accuracy further. Lastly, greedy-backward instead of greedy-forward was also initially considered, but greedy-backward could not compute the optimal set of features in a reasonable amount of time, given our feature space.

As seen from the below results, polynomial kernels generally returned the best results for each classifier type. Also, there was not much difference between the classifiers using band-filtered data and non-band-filtered data. This result actually makes sense, however, as the comparison between the plots of pre-filtered data and post-filtered data showed that not much noise is getting filtered in general.

Table 5. *Accuracy for each classifier using different data sets in an optimal setting*

Classifier	Type	Original	CN+SS*	CN+SS+BF**
1 vs 2,3	Quad	67.4%	74.1%	72.2%
2 vs 1,3	Cubic	79.9%	82.4%	82.4%
3 vs 1,2	Quad	65.2%	67.8%	67.3%
1 vs 2	Quad	67.5%	74.1%	74.1%
1 vs 3	Cubic	67.3%	74.1%	74.1%
2 vs 3	Cubic	66.1%	69.8%	69.8%

*CN+SS: Cepstral normalized with spectral subtraction

**CN+SS+BF: Cepstral normalized with spectral subtraction and band filter

In a general machine learning case, when classifying between 1,2, and 3, the model would take the label that obtains the highest objective value from the set of classifiers, 2,3 vs. 1, 1,2 vs. 3, and 1,3 vs. 2, but since we also have highly functioning single-label vs. single-label classifiers, we made used of them to improve the model's decision making process. Combining all models together, the final results are the following:

Table 6. *Final results from decisions-using-SVM model*

Type	Values
RMS Error	1.13
Classification Accuracy	62.2%

As we can see from the above results, quinary classification accuracy is 62%. This is over 20% improvement from naïve SVM, so the model is performing at a pretty high level, although there is still a room for improvement. Root mean squared error, on the other hand, is not so low considering that best fit linear regression with greedy-forward feature selection will give RMS error of 1.15, but this is because most of the misclassifications occur between 1 and 3. Squared error in this case is 4, so in the worst case, if all our errors come from misclassification between 1 and 3, the result will be 1.2367. Since there can be misclassifications that result in a squared error of 1, our final RMS is mitigated to a value between 1 and 1.23.

4. Future Work

While our results are encouraging, there are many problems that have to be considered before our results can be conclusive. Since dataset we have considered is small, it will be interesting to see what results we can obtain on completely new data. As for PDRS score prediction, although we have found that considering only the end parts of the recordings improves our results, we need more training data to confirm that this is true. As for Hoehn Yahr prediction, the accuracy of the model is decent, but there is still a room for improvement, and we might be able to improve the model further by running analysis on the selected features. In particular, we need to find features to improve the classification between 1 and 3.

Also, although the clean lab data is hard to obtain, we can analyze the clean lab recordings to figure out which frequencies contain predictive power and apply it to our noisy call recordings to emphasize this specific portion of the data.

5. Conclusions

Our goal was to predict proper severities - PDRS and Hoehn Yahr - of Parkinson's Disease patients using their voice recordings. We achieved 62.2% accuracy in predicting Hoehn Yahr score and RMS error of 9.57 in predicting PDRS score. This was possible by minimizing the noise ratio from the original recordings using three different filters: cepstral mean normalization, band filter, and spectral subtraction. Moreover, on top of 38 features, we extracted additional features to make the number of features to be 1,582. Among many machine learning algorithms, support vector machine in decision tree fashion showed the best result over linear regression, decision tree, and k-means clustering. This was possible when we selected effective features by using greedy forward feature selection.

6. Acknowledgements

This study would not have been possible without our project mentors, Jeffrey Ullman, Stanford W. Ascherman Professor of Computer Science (Emeritus), and Dan Jurafsky, Professor of Linguistics and Computer Science at Stanford University. Both Professor Ullman and Professor Jurafsky provided us with valuable intuitions and guidance that shaped the course of this research. We are grateful to Max Little, head of the Parkinson's Disease Initiative and a fellow at the MIT Media Lab, for his expertise on PD classification based on voice recording, and Christine Suver from Synapse.org for providing answers with the dataset. Lastly, we'd like to thank Jun Choi and Bongsoo Suh, Ph.D students in Electrical Engineering for helping us with the implementation of filters, and Sally Park for her support with voice analysis.

7. References

[1] Bayestehtashka, A., and Asgaria, M., and Shafrana, I. “Fully automated assessment of the severity of Parkinson's disease from speech”, Computer Speech and Language, 2013

[2] Bocklet, T., Steidl, S., Noth, E., and Skodda, S., “Automatic Evaluation of Parkinson’s Speech - Acoustic, Prosodic and Voice Related Cues”, INTERSPEECH 2013, 2013.

[3] Bocklet, T., Noth, E., and Stemmer, G., Ruzickova, H., and Rusz, J., “Detection of Persons with Parkinson’s Disease by Acoustic, Vocal, and Prosodic Analysis”,. Online: http://sami.fel.cvut.cz/Articles/Bocklet_ASRU2011.pdf, 2011