IS4250 Report

"Collection and Analysis of a Parkinson Speech Dataset with Multiple Types of Sound Recordings"

Do Thi Huong (A0113779W) – Vu Thi Kim Ngan (A01187613A)

Table of Contents

Background

Our project studies the article “Collection and Analysis of a Parkinson Speech Dataset With Multiple Types of Sound Recordings” by Betul Erdogdu Sakar, M. Erdem Isenkul, C. Okan Sakar, Ahmet Sertbas, Fikret Gurgen, Sakir Delil, Hulya Apaydin, and Olcay Kursun. It was published in IEEE Journal of Biomedical and Health Informatics in July 2013. The IEEE is the largest professional organization advancing innovation and technological greatness for social welfare in the world. IEEE Journal of Biomedical and Health Informatics (J-BHI) publishes original papers interpreting latest advances in the field of biomedical and health informatics where information and communication technologies intersect with health, healthcare, life sciences and biomedicine. This aforementioned article investigates the collected Parkinson dataset using well known machine learning tools like k-nn and svm. Results show that sustained vowels provides more PD-discriminative information. It was also proven that rather than treating each voice sample of each subject as an independent data sample, representing the samples of a subject with central tendency and dispersion metrics enhances generalization of the predictive model.

Parkinson’s Disease

Parkinson's disease (PD) is a neurodegenerative disorder of central nervous system [1]. It is generally observed in elderly people (especially 50 years old and above) and causes disorders in speech and mobility. It is estimated that about 7 to 10 million people are suffering from Parkinson’s disease worldwide [2].

Major symptoms of PD include the following [1]:

- **tremor** of the hands, arms, legs, jaw and face
- **slowness** of movement
- **rigidity** or stiffness of the limbs and trunk
- changes in **speech , voice and swallowing**
- trouble with **handwriting**
- **postural instability** or worsen balance and coordination

PD is a result of the degeneration of a small part of the brain called the substantia nigra. To date, despite of much effort in studies and research, there is no concrete evidence about what is the

cause of such degeneration. Many scientists believe that PD is caused by a combination of genetic and environmental factors, which may be different from person to person [2]. Since the causes remain unknown, there is no absolute cure for the disease. However, appropriate care and treatments are proven to be helpful in improving PD symptoms.

Studies proved that there are three factors that had significant influence on the improvement in condition [3], including:

- the **severity** of the patient's disability before the treatment
- the **duration** that the patient suffers from the disease before the treatment
- the **age** of the patient

This suggests that younger individuals, with a shorter duration of the disease, have a higher chance to get better. However, it can be hard to tell if one has PD. Until now, the best objective testing for PD consists of specialized brain scanning techniques But such tests are known to be invasive and only carried out in specialized imaging centers and can be very expensive [2]. Hence, speech pattern analysis applications of Parkinsonism for building predictive telediagnosis and telemonitoring models become attractive and worth exploring because of two main reasons. Firstly, telediagnosis and telemonitoring systems based on speech signal are cheap and simple to use. Moreover, such systems are believed to aid the development of noninvasive diagnostic for PD.

Data Collection

The data of the study comes from 40 individuals who appealed at the Department of Neurology in Cerrahpasa Faculty of Medicine, Istanbul University. There are 20 People With Parkinson's (PWP) including 6 female and 14 male; and 20 healthy individuals including 10 female and 10 male as shown in Figure 1. 26 voice records were collected from each subject, each voice record includes sustained vowels, words, numbers and short sentences. Eventually, 1040 voice samples were recorded.

Figure 1: Data composition

Methodology

1. Feature extraction

Time-frequency based characteristics are extracted from the voice records using Praat, an acoustic analysis software for the study of speech in phonetics. Praat also assists speech synthesis, including articulatory synthesis. The software is one of the most popular programs among linguists. Using Praat, 26 features were extracted from each voice sample as shown in Figure 2. Information on definition and calculation of the 26 features extracted are provided in Praat manual[1].

[1] http://www.fon.hum.uva.nl/praat/manual/Voice.html

Features	Group
Jitter (local) Jitter (local, absolute) Jitter (rap) Jitter (ppq5) Jitter (ddp)	**Frequency parameters**
Number of pulses Number of periods Mean period Standard dev. of period	**Pulse Parameters**
Shimmer (local) Shimmer (local, dB) Shimmer (apq3) Shimmer (apq5) Shimmer (apq11) Shimmer (dda)	**Amplitude parameters**
Fraction of locally unvoiced frames Number of voice breaks Degree of voice breaks	**Voicing Parameters**
Median pitch Mean pitch Standard deviation Minimum pitch Maximum pitch	**Pitch Parameters**
Autocorrelation Noise-to-Harmonic Harmonic -to-Noise	**Harmonicity Parameters**

Figure 2: Time-Frequency-Based Features Extracted From Voice Samples

2. Classification Methodology

In our research paper, two types of Classification Methodology were carried out including Classification with Leave-One-Subject-Out and Classification with Summarized Leave-one-out.

For Classification with Leave-One-Subject-Out, the researched paper used validation scheme in which all the voice samples of one subject is left out to be used for cross-validation as if it is a testing individual, and the rest of the samples is used for training. According to the LOSO validation scheme, if the majority of the voice samples of a test individual are classified as PWP, then the subject is classified as positive and else negative.

For Classification with Summarized Leave-one-out, the attribute values of 26 voice samples of each subject are summarized using central tendency such as mean, median, trimmed mean (10% and 25% removed) and dispersion metrics such as standard deviation, interquartile range, mean absolute deviation, and a new set of training data consisting of 40 samples of each individual is

formed. Using this methodology, the training data are shortened in sample dimension whereas expanded in feature dimension. The purpose of summarizing the voice samples of subjects is to lower the effect of alterations between different voice samples of a subject.

The two classification methodology was ran using k-nearest neighbor with k value of 1, 3, 5, 7 and SVM classification with linear kernel and radial basis function (RBF)

In the research paper, through evaluating different method by evaluation metric such as accuracy, sensitivity, specificity and MCC, they have achieved the results as follows:

k	Cross Validation	Accuracy (%)	MCC	Sensitivity (%)	Specificity (%)
1	LOSO	53.37	.0007	49.62	57.12
	s-LOO (1-4)	42.50	.0015	30.00	55.00
	s-LOO (2-5)	52.50	.0005	45.00	60.00
	s-LOO (3-6)	50.00	.0000	55.00	45.00
	s-LOO (all)	55.00	.1000	55.00	55.00
3	LOSO	54.04	.0008	53.27	54.81
	s-LOO (1-4)	55.00	.1021	45.00	65.00
	s-LOO (2-5)	60.00	.2010	55.00	65.00
	s-LOO (3-6)	42.50	-.0015	55.00	30.00
	s-LOO (all)	55.00	.1000	55.00	55.00
5	LOSO	54.42	.0009	53.65	55.19
	s-LOO (1-4)	55.00	.1021	45.00	65.00
	s-LOO (2-5)	57.50	.1517	65.00	50.00
	s-LOO (3-6)	50.00	.0000	70.00	30.00
	s-LOO (all)	55.00	.1048	70.00	40.00
7	LOSO	53.94	.0008	54.04	53.85
	s-LOO (1-4)	**65.00**	**.3062**	55.00	75.00
	s-LOO (2-5)	62.50	.2503	60.00	65.00
	s-LOO (3-6)	42.50	-.0017	65.00	20.00
	s-LOO (all)	57.50	.1517	65.00	50.00

Central tendency metrics: 1: mean **2:** median **3:** trimmed mean (25% removed).
Dispersion metrics: 4: standard deviation **5:** mean absolute deviation **6:** interquartile range.

Figure 3: k-NN Classification Accuracies using LOSO and s-LOO

For k-NN Classification, highest accuracy achieved when using s-LOO (1-4) with central tendency of mean and dispersion metric of standard deviation and k = 7. The highest Accuracy is 65% with highest MCC of 0.3062. Result achieved by using LOSO is not desirable as MCC is very near to 0 which mean that the model is not much better than a random guess.

kernel	Cross Validation	Accuracy (%)	MCC	Sensitivity (%)	Specificity (%)
Linear	**LOSO**	52.50	.0006	52.50	52.50
	s-LOO (1-4)	**77.50**	**.5507**	80.00	75.00
	s-LOO (2-5)	70.00	.4082	80.00	60.00
	s-LOO (3-6)	60.00	.2000	65.00	45.00
	s-LOO (all)	67.50	.3504	70.00	65.00
RBF	**LOSO**	55.00	.1005	60.00	50.00
	s-LOO (1-4)	65.00	.3015	60.00	70.00
	s-LOO (2-5)	70.00	.4000	70.00	70.00
	s-LOO (3-6)	72.50	.4506	70.00	75.00
	s-LOO (all)	65.00	.3015	70.00	60.00

Central tendency metrics: 1: mean **2:** median **3:** trimmed mean (25% removed).
Dispersion metrics: 4: standard deviation **5:** mean absolute deviation **6:** interquartile range.

Figure 4: SVM Classification Accuracies Using LOSO and s-LOO

For SVM Classification, highest accuracy is achieved when the with linear kernel model and using s-LOO with central tendency of mean and dispersion metric of standard deviation. The accuracy is higher compare to k-NN Classification at accuracy is 77.5% and MCC is 0.5507.

		LOSO	
		Misclassified	**Correct**
s-LOO	**Misclassified**	8	1
	Correct	11	20
$X_1^2 = 6.75$ α=0.05			

Figure 5: McNemar's Test Between SVM Predictions of LOSO and s-LOO With Mean-Standard Deviation

Comparing across Model, s-LOO is a better model compared to LOSO due to higher accuracy as well as higher number of record being correctly classified.

Classifier	Parameter	Result	Acc. (%)	MCC	Sens. (%)	Spec. (%)
k-NN	1	average	50.61	.0124	52.71	48.52
		best	**82.50**	**.6580**	85.00	80.00
	3	average	49.49	-.0102	54.61	44.37
		best	77.50	.5563	70.00	85.00
	5	average	48.52	-.0298	56.12	40.93
		best	75.00	.5103	65.00	85.00
	7	average	48.12	-.0383	57.18	39.07
		best	77.50	.5563	85.00	70.00
SVM	linear kernel	average	52.06	.0416	54.92	49.22
		best	**85.00**	**.7035**	80.00	90.00
	RBF kernel	average	46.91	-.0618	49.21	44.62
		best	80.00	.6030	85.00	75.00

Figure 6: Average and Best Results of 1000 Runs of Selecting a Random Voice Samples from Each Individual

Comparing across Classifiers of k-NN and SVM, the best model with highest accuracy is SVM with linear kernel.

Hence, from the result that the paper achieved, our group would like to try out with s-LOO method with both k-NN and SVM classifiers.

Classification with Summarized Leave-one-out

1. Data processing

a. Summarized

Since we use Summarized Leave One Out for the methodology, the training data need to be summarized using central tendency such as mean, median, trimmed mean (10%) and dispersion metrics such as standard deviation, interquartile range, mean absolute deviation.

```
mean_sum <- aggregate(train_data[, 2:28], list(train_data$subject_id), mean)
#summarize using mean
sd_sum <- aggregate(train_data[, 2:27], list(train_data$subject_id), sd)
#summarize using standard deviation
median_sum <- aggregate(train_data[, 2:28], list(train_data$subject_id), median)
#summarize using median
mad_sum <- aggregate(train_data[, 2:27], list(train_data$subject_id), mad)
#summarize using mean absolute deviation
meantr_sum <- aggregate(train_data[, 2:28], list(train_data$subject_id), mean, trim = 0.25) #summarize
using trimmed mean (25% removed)
iqr_sum <- aggregate(train_data[, 2:27], list(train_data$subject_id), IQR)
#summarize using interquartile range
```

b. Normalized

In the next process, the data set need to be normalized so that each feature has a 0 mean and standard deviation of 1 or it follows standard normal distribution, the features are fed into SVM and k-NN classifiers for PD diagnosis.

```
# normalized the training data
summ1 <- data.frame(scale(summ[,-54]))
```

2. Classification

a. k-Nearest Neighbor classifier

For the k-NN classifier, we used Euclidean distance metric and k parameter of 1, 3, 5, and 7 and choose k-NN leave one out cross-validation. After getting the prediction results, we use confusion matrix to calculate the accuracy, sensitivity, specificity and MCC of the model.

```
#k-NN
library("kknn")
library("caret")
train <- summ1[,2:53]
cl <- summ1[,54]
out<-knn.cv(train, cl, k = 7, prob = TRUE)
result_knn <- out[0:40]
confusionMatrix(result_knn,cl)
```

b. SVM classifier

For the SVM classifier, we tune the svm using function tune to get the most ideal cost, gamma and epsilon and apply the svm cross-validation to the summarized training data set to achieve the best prediction model. After getting the prediction result, we use confusion matrix to calculate the accuracy as well as sensitivity, specificity and MCC of the model. Then, we apply 2 type of kernel, linear and radial to the cross-validation SVM model to generate prediction.

```
# SVM
library(e1071)
tc <- tune.control(cross = 40)
#linear kernel
obj <- best.tune(svm,class~., data = summ1, tunecontrol = tc, kernel = "linear")
summary(obj)
model <- svm(class~., data=summ1, type = "C-classification", cost = 1, gamma = 0.01886792 , epsilon =
0.1, probability=TRUE)
out<-predict(model,summ1,probability=TRUE)
result_svm_linear <- out[0:40]
confusionMatrix(result_svm_linear,cl)
```

```
#RBF kernel
obj <- best.tune(svm,class~., data = summ1, tunecontrol = tc, kernel = "radial")
summary(obj)
model <- svm(class~., data=summ1, type = "C-classification", cost = 1, gamma = 0.01886792 , epsilon =
0.1, probability=TRUE)
out<-predict(model,summ1,probability=TRUE)
result_svm_rbf <- out[0:40]
confusionMatrix(result_svm_rbf,cl)
```

3. Results and Evaluation Metrics

a. k-Nearest Neighbors

k	Cross Validation	Accuracy (%)	MCC	Sensitivity (%)	Specificity (%)
1	s-LOO (1-4)	47.50%	-0.052	60.00%	35.00%
	s-LOO (2-5)	60.00%	0.201	55.00%	65.00%
	s-LOO (3-6)	52.50%	0.050	50.00%	55.00%
3	s-LOO (1-4)	55.00%	0.102	65.00%	45.00%
	s-LOO (2-5)	67.50%	0.350	70.00%	65.00%
	s-LOO (3-6)	52.50%	0.050	50.00%	55.00%
5	s-LOO (1-4)	55.00%	0.102	65.00%	45.00%
	s-LOO (2-5)	67.50%	0.354	75.00%	60.00%
	s-LOO (3-6)	60.00%	0.218	40.00%	80.00%
7	s-LOO (1-4)	62.50%	0.258	75.00%	50.00%
	s-LOO (2-5)	**72.50%**	**0.451**	**75.00%**	**70.00%**
	s-LOO (3-6)	52.50%	0.056	30.00%	75.00%

Figure 7: k-NN Classification Accuracies using s-LOO with k = 1,3,5,7
Central Tendency metrics: 1:mean 2:median 3:trimmed mean (25% removed)
Dispersion metrics: 4:standard deviation 5:mean absolute deviation 6:interquartile range

We observed that the highest accuracy as well as MCC and sensitivity is obtained while using s-LOO(2-5) summarized by median and mean absolute deviation and k = 7. The result is consistent with the research paper on k = 7 will provide the best model while it is different with the summary factor from the research paper. However, overall we can still observe that the accuracy improved with higher k value, which is consistent with the results of the research paper we based on.

b. SVM

kernel	Cross Validation	Accuracy (%)	MCC	Sensitivity (%)	Specificity (%)
Linear	s-LOO (1-4)	100.00%	1.000	100.00%	100.00%
	s-LOO (2-5)	97.50%	0.951	100.00%	95.00%
	s-LOO (3-6)	97.50%	0.951	100.00%	95.00%
RBF	s-LOO (1-4)	97.50%	0.951	95.00%	100.00%
	s-LOO (2-5)	97.50%	0.951	100.00%	95.00%
	s-LOO (3-6)	97.50%	0.951	100.00%	95.00%

Figure 8: SVM Classification Accuracies using s-LOO with linear and RBF kernel
Central Tendency metrics: 1:mean 2:median 3:trimmed mean (25% removed)
Dispersion metrics: 4:standard deviation 5:mean absolute deviation 6:interquartile range

We observed that the highest accuracy as well as MCC and sensitivity is obtained while using s-LOO(1-4) summarized by mean and standard deviation and using linear kernel SVM which is consistent with the research paper on the best classification method.

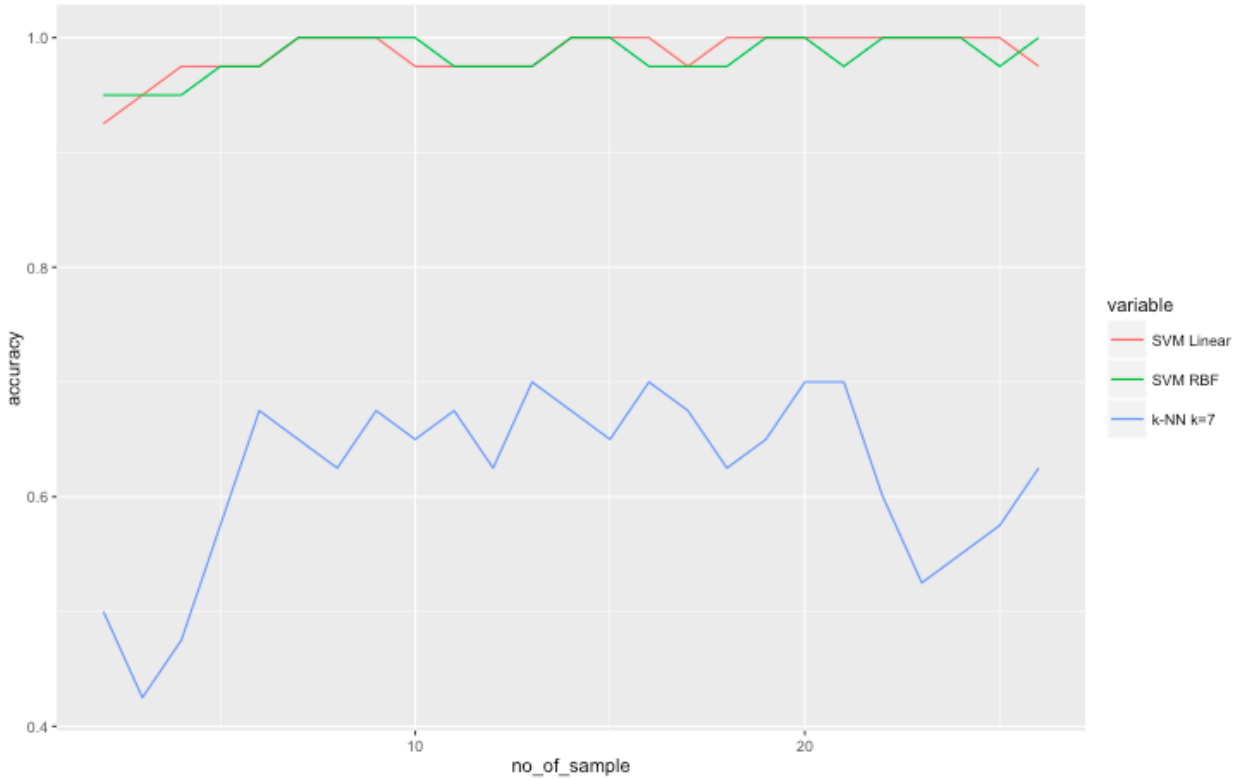
Comparing across the 2 classifiers k-NN and SVM, the best model that help us to obtain highest accuracy, MCC as well as sensitivity and specificity is summarized Leave-one-out with central tendency of mean and dispersion metric of standard deviation using SVM Linear kernel. The result is consistent with the results from our research paper.

4. Further Experimental Results

In order to replicate one of the plot in our research paper, we try the same experiment of feeding each voice sample of the subject individually to the classifier at one time and record the results. We change value of n from 2 to 26 and generate the result for k-NN with k=7, SVM linear and SVM RBF and record in the below table

```
# sample# - corresponding voice samples
# 1: sustained vowel (aaaâ€¦â€¦)
# 2: sustained vowel (oooâ€¦...)
# 3: sustained vowel (uuuâ€¦...)
# 4-13: numbers from 1 to 10
# 14-17: short sentences
# 18-26: words
# Only take n (2-26) first voice sample of each subject_id (starting from 2 to be able to generate standard deviation)
n = 26
train_data <- ddply(train_data, "subject_id", function(z) head(z,n))
```

no_of_sample	knn_k_7	svm_linear	svm_rbf
2	0.500	0.925	0.950
3	0.425	0.950	0.950
4	0.475	0.975	0.950
5	0.575	0.975	0.975
6	0.675	0.975	0.975
7	0.650	1.000	1.000
8	0.625	1.000	1.000
9	0.675	1.000	1.000
10	0.650	0.975	1.000
11	0.675	0.975	0.975
12	0.625	0.975	0.975
13	0.700	0.975	0.975
14	0.675	1.000	1.000
15	0.650	1.000	1.000
16	0.700	1.000	0.975
17	0.675	0.975	0.975
18	0.625	1.000	0.975
19	0.650	1.000	1.000
20	0.700	1.000	1.000
21	0.700	1.000	0.975
22	0.600	1.000	1.000
23	0.525	1.000	1.000
24	0.550	1.000	1.000
25	0.575	1.000	0.975
26	0.625	0.975	1.000



Similar with the experiment in the research paper, our experimental results show that collecting as many voice samples as possible from patients and summarizing the extracted features of each voice sample with central tendency and dispersion metrics increases the success of the diagnostic system. In the above plot, the change in classification accuracy of s-LOO with mean-standard deviation with respect to the increasing number of voice samples is shown. In conclusion, it can be said that using more samples increases both accuracy of the prediction model.

Limitations

As mentioned earlier, voice sample are collected from 40 subjects including 20 PWP and 20 healthy individual. This number of subject is sufficient to apply Central Limit Theorem to the sample. However, in order to get better accuracy, the sampling distribution needs to resemble a normal distribution more closely. It means that more sample points will be required.

Besides, the final data set fitted in the classifiers contain many features (52 or 156 features). This might result in overfitting as number of attributes is greater than the number of observations. Overfitting problem occurs when a model is excessively complex. Such a model will generally

have low predictive performance, as it can overestimate small fluctuations in the data. In order to handle overfitting, there are three approaches that can be adopted, including:

- Getting more observations
- Feature selection
- Tuning of the model

Demographic information and clinical history collected during the study such as age, gender is not taken into account. Studies found out that Parkinson's affects 50 more male than female. About one quarter of PWP have a relative with the disease [4]. These facts prove that demographic and clinical history information will play significant roles in diagnosis of PD.

Conclusion

The study shows a good opportunity to explore the international applicability and validity of the previously built predictive telediagnosis and telemonitoring models, which are presented in the article. Despite of the above limitations, this study has significant contributions in proving that sustained vowels carry more PD-discriminative information than the isolated words and short sentences. Different combinations of central tendency and dispersion metrics (among mean, median, trimmed mean, standard deviation, interquartile range, and mean absolute deviation) is shown to be more useful than using each voice sample of each studied subject as an independent data sample. Summarizing all samples of each subject using mean and standard deviation enhances generalization of the predictive model. As a result, combination of all aforementioned methods has been shown to be a good approach in building such predictive models.

References

[1] Parkinson's Disease: Hope Through Research. (n.d.). Retrieved April 06, 2016, from http://www.ninds.nih.gov/disorders/parkinsons_disease/detail_parkinsons_disease.htm

[2] Statistics on Parkinson's. (n.d.). Retrieved April 06, 2016, from http://www.pdf.org/en/parkinson_statistics

[3] Symptoms of parkinson's disease improve w NADH parkinson's treatment. (n.d.). Retrieved April 06, 2016, from http://nadh.com/pages/parkinsons

[4] Parkinson's Disease (PD). (n.d.). Retrieved April 07, 2016, from https://www.floridahospital.com/parkinsons-disease-pd/statistics