Machine Learning Capstone Project

Suitability of Dysphonia Measurements for Diagnosis of Parkinson's Disease

Project Overview

For my capstone project, I have decided to focus on the field of medicine and classify whether or not a patient has Parkinson's disease based on their vocalization data. For context, Parkinson's is a progressive disease that causes the degeneration of the brain, leading to both motor and cognitive problems. It is thus reasonable to assume a correlation between a patient's ability to speak and their progression into Parkinson's as these capabilities regress. The data set I worked with was obtained through a 2008 study by the journal, *IEEE Transactions on Biomedical Engineering*, of how various parameters of voice frequency can help classify if a patient is suffering from Parkinson's. By performing a classification on this data, I hope to prove that vocalization tests are indeed a well suited way to diagnose a patient for this disease. The diagnosis at hand is valuable as the test can be administered at home without the need for the patient to visit a hospital – saving time, cost, and energy.

Problem Statement

This project attempts to prove that vocalization data from a patient can help diagnose whether or not they suffer from Parkinson's. As such, it is initially assumed that there is a relationship between the two. I will attempt to run various machine learning classifiers (Ex. Naïve Bayes, SVM, etc.) on the data in hopes to reach a high predictability rate that is matched with a reasonable runtime. The study itself obtained a predictability rate of 91.4% and so I hope to reach a rate close to this or possibly to surpass it. If I can reach a rate that is within a 5% interval of the one obtained in the study, I will have proved the study correct.

Metrics

I will be using an F1 score to measure the predictability rate of my classifier. This is a reasonable way to evaluate my model's performance as it tests the accuracy by considering both precision and sensitivity. This ensures that a high score will produce the least amount of false positives and negatives, which is important as we are trying to accurately classify if a patient has Parkinson's. Frequent misdiagnosis would be daunting for them from this form of analysis and also make vocalization ineffective. This is why the F1 score is a good value to follow as I try to optimize my model.

Data Exploration

I have thoroughly analyzed the data set and come up with these conclusions. There are a total of 195 rows of data or patients with 23 feature columns. The voice measurements are from 31 people, 23 with Parkinson's disease and 8 without. In this data set, 147 of the rows have Parkinson's whereas 48 do not. The features themselves are related to the frequency if a patient's vocalizations. The attributes are listed below:

Name - ASCII subject name and recording number

MDVP:Fo(Hz) - Average vocal fundamental frequency

MDVP:Fhi(Hz) - Maximum vocal fundamental frequency

MDVP:Flo(Hz) - Minimum vocal fundamental frequency

MDVP:Jitter(%),MDVP:Jitter(Abs),MDVP:RAP,MDVP:PPQ,Jitter:DDP - Several measures of variation in fundamental frequency

MDVP:Shimmer,MDVP:Shimmer(dB),Shimmer:APQ3,Shimmer:APQ5,MDVP:APQ,Shimmer:DDA - Several measures of variation in amplitude

NHR,HNR - Two measures of ratio of noise to tonal components in the voice

status - Health status of the subject (one) - Parkinson's, (zero) - healthy

RPDE,D2 - Two nonlinear dynamical complexity measures

DFA - Signal fractal scaling exponent

spread1,spread2,PPE - Three nonlinear measures of fundamental frequency variation

A data sample of both a healthy patient and one with Parkinsons is presented as well:

name	MDVP:Fo(Hz)	MDVP:Fhi(Hz)	MDVP:Flo(Hz)	MDVP:Jitter(%)	MDVP:Jitter(Abs)
phon_R01_S06_6	162.568	198.346	77.63	0.00502	0.00003
phon_R01_S07_1	197.076	206.896	192.055	0.00289	0.00001
MDVP:RAP	**MDVP:PPQ**	**Jitter:DDP**	**MDVP:Shimmer**	**MDVP:Shimmer(dB)**	**Shimmer:APQ3**
0.0028	0.00253	0.00841	0.01791	0.168	0.00793
0.00166	0.00168	0.00498	0.01098	0.097	0.00563
Shimmer:APQ5	**MDVP:APQ**	**Shimmer:DDA**	**NHR**	**HNR**	**status**
0.01057	0.01799	0.0238	0.0117	25.678	1
0.0068	0.00802	0.01689	0.00339	26.775	0
RPDE	**DFA**	**spread1**	**spread2**	**D2**	**PPE**
0.427785	0.723797	-6.635729	0.209866	1.957961	0.135242
0.422229	0.741367	-7.3483	0.177551	1.743867	0.085569

There seems to be no abnormalities about the input space or data set to report. No values are missing from the data set and no outliers have been identified when the data was graphed and analyzed.

Exploratory Visualization

To get a better grasp of the data, I developed a scatter matrix (Appendix Figure 1.0) that allows me to see the relationship between the features presented from the data set. There are multiple correlations seen in the matrix, but I specifically analyzed the distribution of a feature when compared to the status of the patient. I noticed that across all features, those diagnosed with Parkinson's had a larger spread of data than those who did not. While healthy patients had their data clustered near 0, the Parkinson patients had the features run across the entire axis. This

results in a separation between the status of 1 and 0 in the spread1, spread 2, D2, and PPE columns. This division will allow for greater success by SVM classification, which can better separate the data with the hyper plane.

Algorithms and Techniques

I have implemented multiple supervised learning algorithms to analyze the data set. My hopes are to find the classifier that is most optimal for this problem. Afterwards, I will be able to tune it to optimize its performance. The end goal is to optimize a model off of one of these classifiers so that it can perform as well or better than the predictability rate of those who wrote the study. I have chosen three different supervised classifiers as listed below:

Naïve Bayes: A classifier that makes use of Baye's theorem to try and classify data. The use of Baye's equation makes use of the likelihood of a specific classification based on probabilities returned from the features of the data. I decided to implement this classifier as it is known to only require a small training set to estimate parameters and is not sensitive to irrelevant features.

Support Vector Machines: A classifier that categorizes the data set by setting an optimal hyper plane between data. I chose this classifier as it is incredibly versatile in the number of different kernelling functions that can be applied. I hopes are that when tuned, this model can yield a high predictability rate.

Stochastic Gradient Descent: A classifier that applies the concept of gradient descent to reach an optimal predication rate. I chose this classifier because it has multiple parameters that could potentially be tuned if it produces strong results under my data set. My main concern is that it tends to work well with sparse data, but is sensitive to larger training sets.

Gradient Tree Boosting: An ensemble classifier (combines predictions of base estimators) that makes use of multiple decision trees - giving their contributions weights – to classify data. The weights are developed using the gradient descent algorithm. I decided to implement the GTB because of its strength in handling heterogeneous features and high accuracy. The problem I do expect with this classifier is that it scales poorly and so with larger sets of data, this would not be optimal. It is still interesting to see the results of this classifier however.

Benchmark

The benchmark for this project is the resulting prediction rate from the original study. I expect to obtain a predictability rating that is at least 5% within the margin of the 91.4% they obtained in their analysis to be successful.

Data Preprocessing

There was no need to implement feature transformation, as all of the data is continuous. The data itself does not project any outliers that I feel would have an enormous impact on the classifiers I have created. Thus no effort has been made to seek out and eliminate data points.

Implementation

The supervised learning algorithms ran smoothly with the given data set. I encountered no issues with the metrics and techniques I applied. Some effort was made to tinker with the scatter matrix I produced so that the labels were clearly visible. I should report that I have created multiple functions to help in the training and prediction from my classifiers. This makes it relatively easy to do so for multiple classifiers as I am.

Refinement

I successfully produce F1 scores based off the predictions made by my classifiers. Below I have listed the results in a table (NOTE: Time is in seconds):

Naïve Bayes

Training Set Size	Prediction Time (Train)	Prediction Time (Test)	F1 Score (Train)	F1 Score (Test)
50 (~25%)	0.0003	0.0002	0.8406	0.8615
100 (~50%)	0.0002	0.0003	0.7967	0.8000
147 (~75%)	0.0003	0.0002	0.7650	0.8000

Support Vector Machines

Training Set Size	Prediction Time (Train)	Prediction Time (Test)	F1 Score (Train)	F1 Score (Test)
50 (~25%)	0.0005	0.0004	1.0000	0.8780
100 (~50%)	0.0007	0.0012	0.9931	0.8780
150 (~75%)	0.0012	0.0004	0.9867	0.8675

Stochastic Gradient Descent

Training Set Size	Prediction Time (Train)	Prediction Time (Test)	F1 Score (Train)	F1 Score (Test)
50 (~25%)	0.0002	0.0002	0.8764	0.8571
100 (~50%)	0.0009	0.0002	0.5789	0.5660
150 (~75%)	0.0003	0.0004	0.0177	0.0526

Gradient Tree Boosting

Training Set Size	Prediction Time (Train)	Prediction Time (Test)	F1 Score (Train)	F1 Score (Test)
50 (~25%)	0.0004	0.0003	1.0000	0.8974
100 (~50%)	0.0004	0.0003	1.0000	0.9600
150 (~75%)	0.0006	0.0007	1.0000	0.9600

The data set shows that the Support Vector Machine and Gradient Tree Boosting classifiers yielded the highest and most consistent in their F1 score. From here I decided to select SVM as my classifier to tune. The reason why is that while GTB was successful, it does not scale well with larger sets of data and so with larger data sets, its accuracy would dramatically drop.

Model Evaluation and Validation

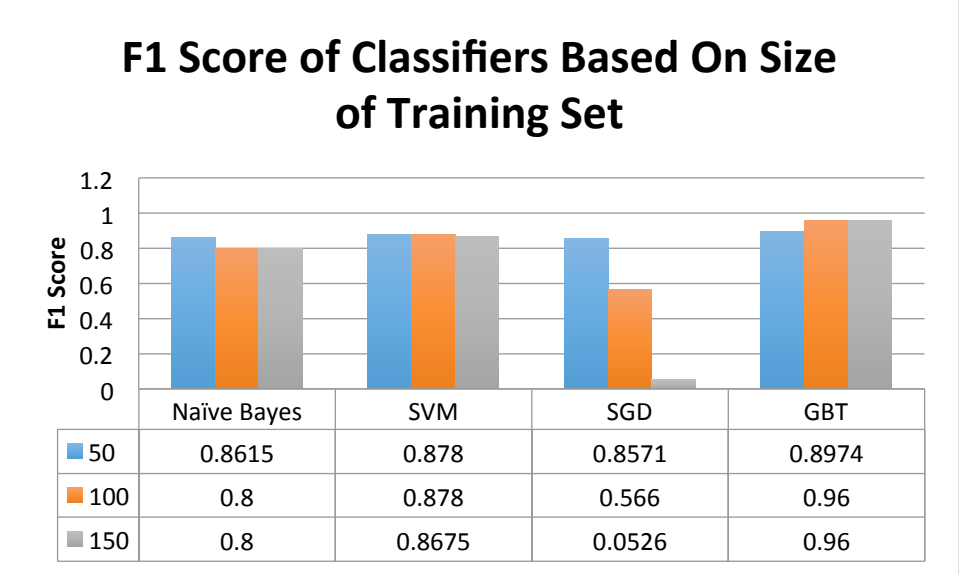
The final model aligns with the solution expectations. It produces a F1 score under the best parameters from tuning. I have tested the model under unseen data I have created and it works

successfully. The SVM classifier itself is inherently robust to small changes in the training data. We can trust the results from the model, but in industry it may not be so. This is because the final accuracy rate of the SVM was 89.74%, which has a large error for the healthcare industry. Most diagnosis methods must be 97% accurate for them to be applied on patients.

Justification

The final results are not as strong as the benchmark provided by the study of 91.4%. My model's final rate was 89.74%. However, this does verify the study as my model is within a 5% margin of their analysis. I believe with even further scrutiny in the tuning of the model, I may be able to achieve the same result as that with the study. It may be possible to achieve even a higher rate but that would require a very thorough optimization of the parameters and cleaning of the data for small outliers.

Free-Form Visualization



	Naïve Bayes	SVM	SGD	GBT
50	0.8615	0.878	0.8571	0.8974
100	0.8	0.878	0.566	0.96
150	0.8	0.8675	0.0526	0.96

For my free-form visualization, I have graphed the F1 scores of the four classifiers I have tested. The two classifiers that were put under scrutiny by me were the SGD and GBT. This is because one of them had a dramatic drop in its F1 score while the other had a higher rate than that found

in the study (91.4%). With more research into both classifiers I was able to identify why this was the case. The SGD classifier is better when data is sparse (ex. Text classification), but is sensitive otherwise. The GBT classifier looks as if it is excellent, but underperforms as the amount of data scales up. This means that it would underperform when there is much more data and so it would not be viable in industry.

Reflection

My project went through the analysis of multiple classifiers in hopes of finding one that would verify the studies classification using vocalization data. I found that the spread of the data for those inflicted with Parkinson's over those who were healthy to be very interesting. This indeed supports the fact that vocalization data correlates with whether or not someone has Parkinson's. The difficult task I encountered was coding and analyzing the scatter matrix to notice this relationship. However, it was very helpful in allowing me to better understand the data set. This model fits my expectations as it verified the study. Unfortunately, as it does not reach the 95% confidence interval, I do not believe it maye be used in the general setting of healthcare for diagnosis.

Improvement

The model itself can be further tuned to try and improve the prediction rate. I was able to figure out how to use all the algorithms I implemented, but I know that more supervised learning classifiers exist and they have potential in classifying this data. As for my final solution, I believe a better solution exists with more time implemented to optimizing the parameters of the classifier and analyzing more supervised classifiers. A classifier that is created for specifically this type of problem could perform the best.

Bibliography

Parkinson's Dataset Study: IEEE Journal of Biomedical and Health Informatics. (n.d.). Retrieved June 02, 2016, from http://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=6221020

Supervised Learning Explanation and Strength/Weaknesses: Scikit-learn. (n.d.). Retrieved June 02, 2016, from http://scikit-learn.org/stable/

Appendix

Figure 1.0 – Scatter Matrix of Features From Parkinson's Data Set. Expand image to see details.