Identifying Parkinson's Disease from Passively Collected Data

Michelle Wang
mjwang@mit.edu
MJFF Challenge
25 March 2012

1. Introduction

The increasing prevalence of smartphones and smart computing devices has created new opportunities for passive data collection and analysis. These commercial devices that we bring around every day carry a variety of sensors that provide a multitude of motion, environment, and position data that can be used to infer information about the user and their surroundings.

The goal of this project was to use passively collected acceleration data to monitor and identify patients with Parkinson's disease and look for broader applications of the techniques used. By processing the collected data to extract meaningful features and applying machine learning techniques to classify the data set, it was shown that patients with Parkinson's disease could be identified with a high rate of reliability. This is a promising result as acceleration data is available and applicable in many different technologies, and the techniques used can easily be extended and applied in different devices and settings.

2. Data

The available data consisted of acceleration, audio, battery, compass, and GPS information collected from an Android smartphone placed in the subject's pocket or worn around the neck. I chose to focus purely on the acceleration data in searching for Parkinson's indicators as it was likely to reveal the most information about a user's motion and behavior.

Acceleration data was recorded approximately every two seconds for periods of many hours, providing an extremely large amount of densely packed points. This made it difficult to find overarching trends and visualize a bigger picture. To combat this problem, the acceleration data was decimated such that values were averaged over hourly windows rather than the original two second window.

An important point of variability in the acceleration data was that the x,y, and z axis were measured with respect to the phone's orientation. Thus, the x,y,z axis could vary from person to person depending on the position of the phone and how it was carried around. To eliminate this variability, the x,y,y channels were combined into one by taking the root mean square. Several other ways of achieving this were also explored including taking the maximum, minimum, or sum of the x,y,z values.

After reducing the granularity of the provided data and eliminating x,y,z variability, the data was much easier to visualize and understand. I chose to investigate the raw acceleration values along with acceleration data from the four power spectral density channels generated from the original data. This provided five separate data features.

Different pairs of these five features show clear trends and can be seen plotted in Appendix 1, Figure 1.

3. Methods and Results

With the newly processed data showing promise, I chose to use a Support Vector Machine (SVM) to generate linear and radial basis classifiers that could distinguish between Parkinson's patients and control subjects. The Support Vector Machine was implemented using the LIBSVM library in MATLAB.

The data was separated into five training groups with around 800 data points per group. The training accuracy across all five groups using a linear separator averaged ~86.6% while the radial basis classifier yielded accuracies of ~99%.

K-fold cross validation was used to verify the classifier's abilities to predict correct outcomes from new data. The data was separated once again into five groups with four of the five group used for training and the remaining group left as validation data. Using a linear classifier, the algorithm was able to correctly distinguish Parkinson's patients from control subjects with only ~15% error on new data. A radial basis classifier returned even better results of ~9% error, which is commensurate with the theoretical expected generalization error estimate.

These results show that the possibility of identifying Parkinson's patients from passively collected data is extremely feasible. Other aspects of classification to examine include measuring the trade-off between computational complexity and accuracy. The radial basis kernel performs better, but it is unclear whether benefit of the computation outweighs the cost as the linear kernel does relatively well also. If these algorithms are to be run on the device, power consumption may be an additional constraint.

For illustration purposes, a visualization of a radial basis SVM classifier separating a smaller subset of the acceleration data using the low power spectral density and low-mid power spectral density features can be seen in Appendix 1, Figure 2. The actual classifier was trained and validated in five dimensional space using all features.

5. Further Applications

While smartphones are convenient, widely available devices that contain the necessary sensors for Parkinson's analysis, the methods above can potentially be extended to future medical devices with embedded accelerometers. Data from these devices might be collected and processed in real time with built-in algorithms on the device or examined after the data has been exported.

In addition, expanding the classifiers above to identify discrete stages in disease progression rather than a simple binary classification has the potential to create the basis for a chronic diagnostic tool for Parkinson's therapies or other movement disorders. With further work, the data from the accelerometers might one day be processed, classified, and used by clinicians as an additional input to guide treatment.

6. Conclusion

In summary, the objective of these experiments was to develop a method for using passively collected data to monitor Parkinson's patients and distinguish them from control subjects. After processing the data to eliminate x,y,z variability and extract meaningful features, a Support Vector Machine was able to identify the Parkinson's patients from control subjects with excellent accuracy. This classification idea can be extended to monitor disease progression by labeling patients based on a disease state rather than just 'PD' or 'Non-PD'. This may require additional machine learning tools and feature extraction. However, the preliminary results for binary classification are promising and expansion is worth exploring. Furthermore, if successful, these algorithms can potentially be used in implantable medical devices and tied into chronic treatment systems for Parkinson's disease.

A1. Figures and Plots

Figure 1. 2D Combinations of Different Acceleration Features

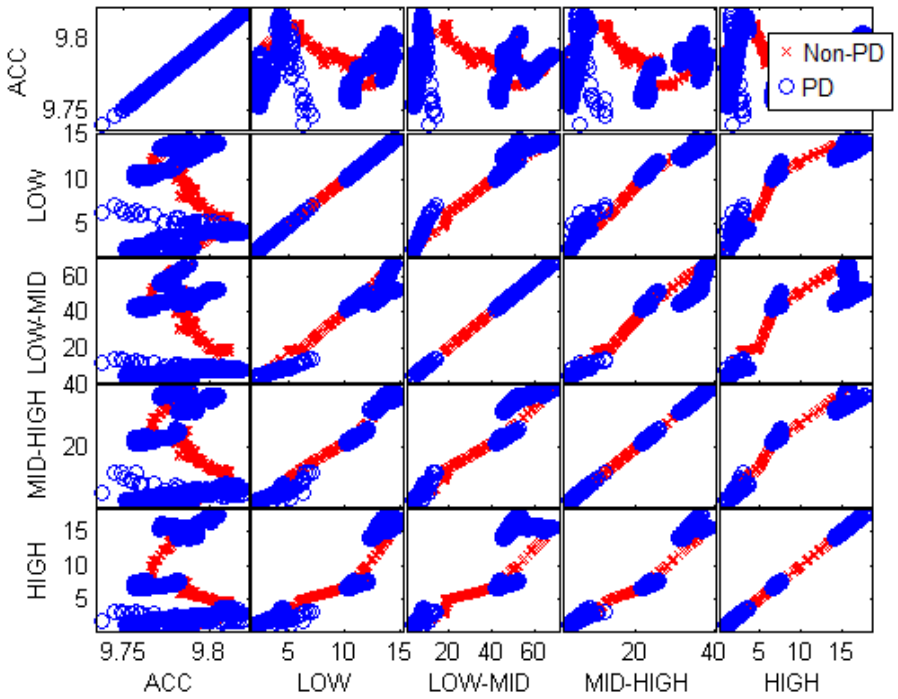
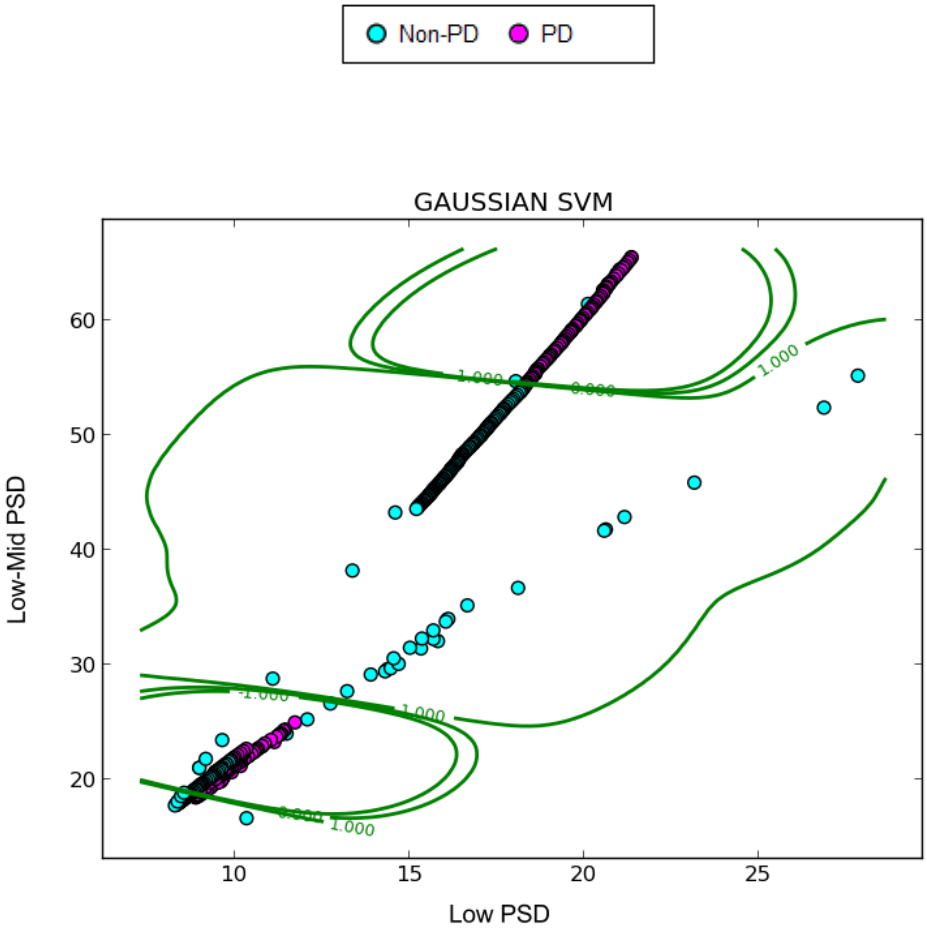


Figure 2. 2D Radial Basis SVM Classification with Low PSD and Low-Mid PSD Features



A2. Team and Contact Information

Michelle Wang
mjwang@mit.edu
763-447-8607
MIT Class of 2013
Department of Electrical Engineering and Computer Science