Supervised and unsupervised machine learning for the detection, monitoring and management of Parkinson's disease from passive mobile phone data.

Submitted by LIONSolver, Inc. (M. Brunato, R.Battiti, D. Pruitt, E. Sartori)

Executive summary

We demonstrate that a Machine Learning approach is superior to conventional statistical methods for the detection, monitoring and management of Parkinson's disease. In spite of the very sparse data of this specific Parkinson's diagnosis problem, we can predict incidence and monitor progression of the disease with 100% accuracy on the competition data. In addition to producing accurate detection, a machine learning approach paves the way for disruptive innovation in the monitoring and management of the disease, given the following advantages associated with this branch of artificial intelligence:

- Discovery of *hidden and novel relationships* through unsupervised learning (clustering), leading to the identification of different Parkinson's disease variations and comparative effectiveness of treatments.
- Construction of robust models drawn from abundant data that will resist data errors and *improve as more data about patients are collected*. This is a similar approach taken by IBM Watson in cancer therapy support at the Memorial Sloan-Kettering center.
- Creation of an intelligent mobile application that *gamifies* patient input to augment passively collected data and provides interactive feedback to patients and monitoring summaries for physicians.
- Automated ranking of different inputs (attributes) by measuring their *information content*. This produces a Parkinson's management system for patients and physicians that continuously 'learns from data.'
- Powerful visualization of similarity among patients though dimensionality reduction (*similarity maps*) that allow people to rapidly assess progress of the disease relative to the population as a whole.
- Scalability to very large number of patients through parallel and distributed computing (*cloud*) that would be useful in producing population-wide understanding of disease progression, and the comparative effectiveness of therapies.
- The possibility to aid drug development in the pharmaceutical sector by combining machine learning with optimization [3] to understand commercial patient-wide effectiveness of approved medications for Parkinson's sufferers.

In spite of the limited set of available patients' data, we demonstrate that our approach to measuring similarity among patients is a building block for creating a robust classification of all patients as a population through the combination of multiple diagnostic classification systems (including more traditional statistical ones) by "democratic" techniques [2]. The preliminary results about diagnosis, clustering, similarity maps, and attribute selection shown in this exercise offer promising potential, and data from a larger patient sample will permit us to validate this hypothesis.

In the following sections, we address three major objectives of this competition around detection, progression of the disease and creative ways of using the data to improve patient quality of life, while validating the novelty of our approach (the fourth competition objective).

Additional detail regarding our approach is included in the Appendix.

Objective 1: Distinguishing Parkinson's sufferers from the broader participant group.

Because data are collected in a passive way, the movement-related symptoms of Parkinson's (shaking, rigidity, slowness of movement) are overshadowed by other effects of normal mobile phone movement that must be considered as "noise" sources (example: walking or riding in a moving car), and they need to be filtered out to make the "signal" emerge.

Instead of proceeding by hand, we use an unsupervised machine learning (clustering) approach: in addition to being extremely accurate, it automates cleaning, increases speed, and offers the flexibility to adapt to different conditions. This would be very important in larger scale versions of this experiment.

Figure 1 describes the procedure. Data from each 1-hour recording session have been clustered with respect to the compass readings (pitch, roll and azimuth, describing the inclination of the device), the GPS data (longitude and latitude) and recording time. A dense cluster represents a short time window during which the subject wasn't moving (white bands in Figure 1). Next, the data from each time window is aggregated into the mean value and standard deviation of the power spectrum density at various frequencies, providing information about the presence of tremors during a time window.

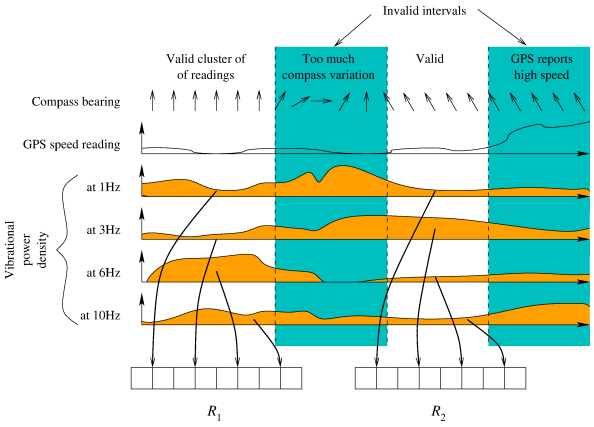

Figure 1 - Accelerometer data clustering by removal of time windows when external movement is detected.

Each valid time window is therefore summarized into an 8-dimensional vector of mean values and standard deviations for the power spectral density at the frequencies of 1Hz, 3Hz, 6Hz, and 10Hz:

$$R = (m_{1\text{Hz}}, e_{1\text{Hz}}, m_{3\text{Hz}}, e_{3\text{Hz}}, m_{6\text{Hz}}, e_{6\text{Hz}}, m_{10\text{Hz}}, e_{10\text{Hz}})$$

A projection of these data point in two coordinates, $m_{3\text{Hz}}$ and $m_{10\text{Hz}}$, is provided in Figure 2 where Parkinson Disease (PD) dots and Control Group dots (CG) tend to occupy different positions on the plot, suggesting that different mixtures of frequencies can be measured in the two groups, in particular - as expected from the medical literature - in the 4-6Hz frequency band typical for Parkinson-related tremors.

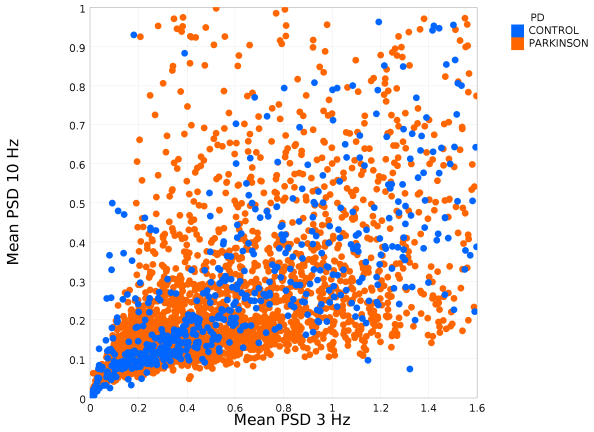

Figure 2 - Power spectral density at 3Hz and 10Hz at each valid time window for Parkinson and Control subjects

Scatter-of-points Characterization

The existence of local correlations between measurements at different frequencies, particularly visible after noise has been removed, motivates the adoption of Machine Learning techniques as a robust approach to PD subject classification that relies on the automatic detection of significant attributes (please refer to the Technical Appendix for details). As explained above, every test subject is associated to a "cloud" of points containing information about accelerometer powers at different frequencies and at different times. A measure of the overlap of different subjects' clouds is thus a convenient input to a supervised machine learning algorithm. As a simplified example, consider three subjects whose point clouds look like those depicted in Figure 3. Subjects 2 and 3 are qualitatively similar, and possibly belong to the same category, because their point clouds have a large overlap. On the other hand, subject 1 has a comparably different behavior, and belongs presumably to a different category. In order to compute the overlap of point clouds, we need to overcome three problems: *variability* (different subjects have a different number of points, and the cloud overlap must also depend on point density), *dimensionality* (the clouds extend in possibly many dimensions), and the need for a *fixed representation* (in order to employ a machine learning algorithm, clouds must be conveniently represented by a fixed number of parameters called *attributes* (or *features*).

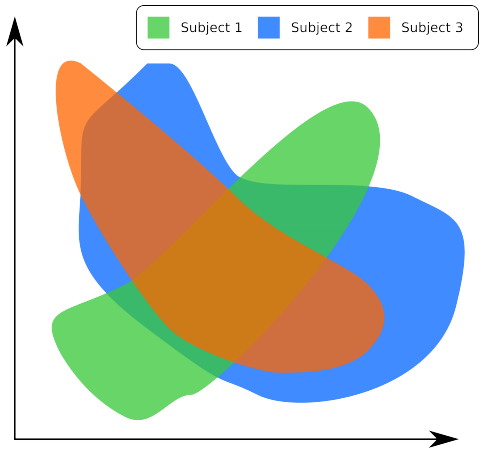

Figure 3 - Overlapping clouds for different test subjects.

We chose the *quantization* approach: the whole space is divided into a convenient number of subintervals, and every cloud is characterized by the number of points contained in each subinterval. The dataset from every patient is therefore reduced to a fixed sequence of numbers, and these sequences can be directly compared to each other by using algorithms.

The Support Vector Machine algorithm [4,5] typically performs well even for small datasets and sparse attribute vectors. The algorithm has been coupled with a Radial Basis Function kernel and tested by leave-one-out cross-validation (each sample, in turn, is treated as a test instance, while all the others are used to train the system) with our proprietary LIONsolver [6] software.

Figure 4 reports the conclusions. Using Machine Learning produced results compatible with a 100% accuracy (16 correct classifications out of 16 test cases) for a wide range of parameter values, even when just one tremor frequency is used for generating the attribute values. In spite of the small number of samples, the system displays a good generalization performance (performance for patients not considered during the training phase).

	d=1		*d=3*	
	Q=2	*Q=5*	*Q=2*	*Q=5*
# features	2	5	3	13
$\gamma = .001$	15	14	**16**	15
$\gamma = .01$	**16**	**16**	**16**	**16**
$\gamma = .1$	**16**	**16**	**16**	**16**
$\gamma = 1$	**16**	**16**	**16**	**16**
$\gamma = 10$	**16**	**16**	**16**	**16**

Figure 4 - Number of correct leave-one-out classifications over 16 samples for different training parameter values.

This automated, rapid-learning approach would be enormously powerful for analyzing large patient populations and for monitoring tremors across a great range of frequencies (including very noisy environments).

Objective 2: Measuring the progression, change, and variability of symptoms in PD subjects.

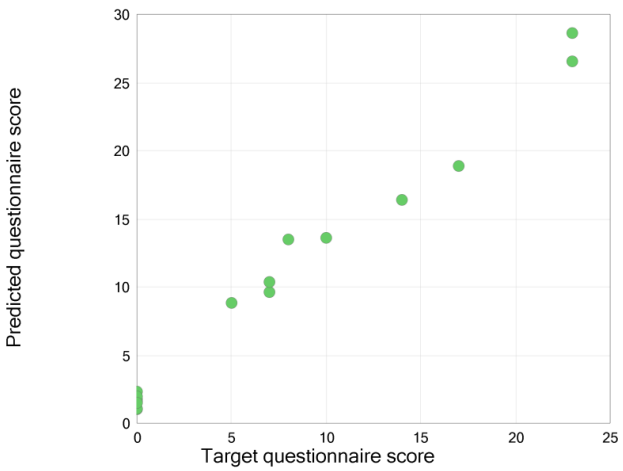

Figure 5 - Predicted severity versus actual severity as measured by scoring questionnaire answers.

To be really useful, a Parkinson's management approach cannot just output the "yes/no" outcome of a binary classification task. In particular, more significant outputs include scoring that associates severity with each subject.

In particular, the ability to provide an objective score as the output of the passive data analysis is important for monitoring the progression of the disease over subsequent months and years. In order to validate our approach to the problem, the severity assessment questionnaire included in the competition data provides a numeric (the sum of scores from specific questions about symptoms) that can be used as a severity value for the disease.

A linear regression model based on the same attribute sets as in the binary classification problem described earlier yields the results shown in Figure 5, where the predicted questionnaire scores are plotted against those observed for the 16 test subjects. The Figure suggests a strong correlation between the accelerometer measurements, used as inputs, and the questionnaire score, although tests are not conclusive because of the small sample set.

The progression of the disease can also be visually examined by means of a *Similarity Map*, where subjects are represented as dots in a two-dimensional plane and are plotted in such a way that similar subjects tend to be close to each other. For this analysis (see Figure 6) *similarity* has been obtained by measuring how close the point clouds of two subjects are in the Gamma parametric space (see Appendix E for details).

We can observe that the control subject APPLE is very close to the group of PD patients: this may be due to his age, which may result in tremors similar to the ones caused by Parkinson disease. Subject CHERRY, on the other hand, is a Parkinson's sufferer, but his/her form of disease seems to be in a less-advanced stage, considering the information in the questionnaire. We can expect that, as the disease worsens, or improves due to new treatments, a Parkinson's

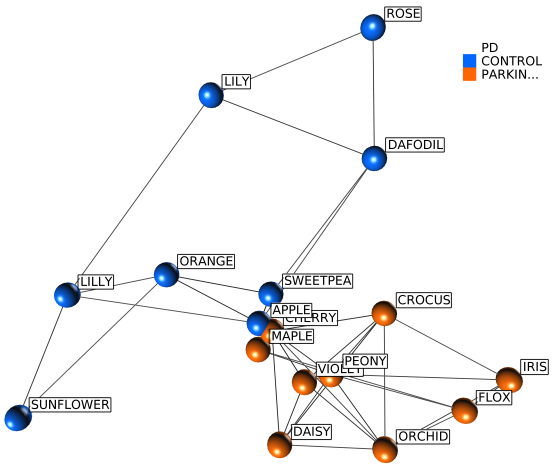

Figure 6 - Similarity map for Parkinson and Control patients based on statistical features.

subject's dot will "drift" from the center of the map towards the lower quadrant or vice versa.

Being able to correlate the severity of the disease (as measured by the questionnaire) to the collected passive data is fundamental to *track the progression of the disease* and to offer physicians a valuable, objective and independent methodology to measure progress related to treatment. In turn, being able to correlate progress to treatment in an objective way will enable researchers and physicians to promptly identify anomalous conditions related to inefficient therapies or non-collaborative patients. To this aim, systematic data collection is crucial and the long-term mobile phone information must be complemented by data about drug assumption and eventual therapy suspension periods. The active monitoring approach described below is important in keeping the patient involved and collaborative during the data collection process.

Objective 3: Using the data in creative ways to inform patient care and improve their quality of life.

Passive and Active Monitoring

Parkinson's disease symptoms, such as tremors, can be triggered by some simple activity requiring concentration and in all cases have an impact on dexterity. We built a web-based, mobile application using game mechanics to actively monitor some simple activities and provide feedback to the patient and to their clinician. We developed a simple set of exercises and coded the prototype to accept inputs from mobile phone usage. While the user interacts with various tasks, the phone stores and later sends data to a cloud-based system for aggregation and delivery to a physician, including web-based dashboard visualizations that would provide information on the patient's disease progression. As this is a prototype service, we have not fully considered the ways of integrating this data into physician workflow.

The first exercise asks the user to hold the phone in a hand for ten seconds, then to switch to the other hand. The later phases of the exercise require the user to hold the phone while completing a simple task (composing a word by using two large buttons).

The main idea of the game is to offer a simple test of how concentration and dexterity are triggering tremors, and to capture their extent and change over time. One potential feature of this application is that a patient doesn't need to download and install an app for that: the phone browser is powerful enough to collect and submit the data (it works on iPhone iOS 5 and above, Android 4.0 and above).

Depending on user studies regarding patient willingness and engagement, existing passive data could be augmented with other passive data collected by a custom app including voice quality and intensity recording, sleep motion and volume sensors to measure sleep onset apnea, insomnia, sleep fragmentation and REM disturbances. Patterns and suspicious interruptions in the gathered data can be analyzed to find known patterns, while new patterns can be identified by Machine Learning techniques applied to existing data, and novelty detection algorithms can alert about yet unknown conditions. Active voice recordings can also help assessing the quality of the patient's voice by detecting known speech patterns and learning new ones. This whole-life view for a Parkinson's patient and for their clinicians might reveal lifestyle, medication and other causal factors impacting patient quality of life.

Figure 7 - A sample session record, as visualized on the physician's browser.

Data from all monitoring is captured and logged for each patient and retained in a secure, cloud-based repository.

Figure 7 shows a sample session recording, being displayed on the physician's web application. It displays, the frequency spectrum of the accelerometer readings (light red means that at a given time little energy was concentrated at the given frequency, darker red means that more tremors were evident), together with the relevant actions of the test subject, represented as correct or wrong responses to the given tasks, and the compass readings, useful to check whether the patient was really still while performing the exercise.

We believe that this application can effectively support physicians by providing a way to monitor the development of their patients' disease with much higher frequency than the handful of visits per year that are possible now.

Moreover, the collected data can easily be decoupled from identity information, so that anonymized datasets can be collected for population-wide status and progression. This has potential implications for research and

development, as it would provide summary data on comparative effectiveness of therapies and offer researchers an insight into progress of therapeutic development.

Objective 4: Do the analyses and proposed uses of the data use innovative approaches and methods?

We have demonstrated that a machine learning approach appropriate for the very sparse data of this specific Parkinson's diagnosis problem, with a limited set of patients and the goal of scaling to millions of Parkison's sufferers offers a superior approach to other methodologies. The Machine Learning approach, in turn, opens up the possibilities of other innovative approaches related to scalability when the data population grows. In this Section we discuss techniques for reducing the size of the training sets by means of attribute set reduction (feature selection) and for speeding up the computation via parallelization and cloud computing.

Scalability: parallelization and cloud computing

Additional patients are needed for a more precise validation of the proposed techniques. The advantages offered by the availability of larger samples are evident, and range from better noise reduction and identification of borderline cases to data-driven assessment of drugs efficacy and treatment protocols. The proposed processing techniques are highly parallelizable, and can thus take advantage of modern cloud computing resources. Preliminary experiments on 300 computers hosted on Amazon's EC^2 platform show that the most critical steps can be implemented in a suitable platform with a speedup close to the theoretical limit. See Appendix D for more detail.

Scalability: Attribute selection

The identification of an appropriate set of inputs is crucial both to ensure the best possible generalization results and to identify relevant "signals" for relating inputs to outputs, in particular for *complex nonlinear tasks*. The *mutual information* (MI) criterion can be used to evaluate a set of candidate attributes and to select an informative subset. Because the MI measures arbitrary dependencies between random variables, it is suitable for assessing the "information content" of attributes in complex classification tasks, where methods based on linear relations (like the correlation) are prone to mistakes. An algorithm based on a "greedy" selection of the attributes and that takes both the MI with respect to the output class and with respect to the already-selected attributes into account has been proposed in [1], and it has recently being applied to a growing number of problems in bioinformatics and health care.

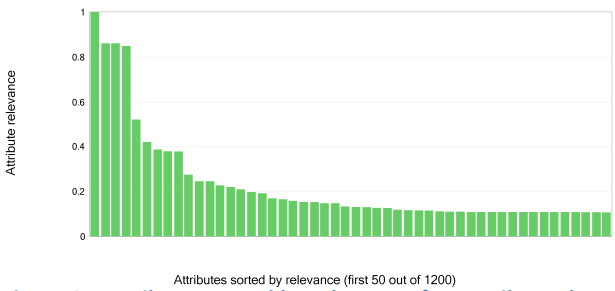

Figure 8 - Attributes sorted by relevance: few attributes have high impact.

Because of the very limited number of patients is not sufficient for a full test of MI, a preliminary distribution of relative attribute weights is obtained for a first-order regression task for questionnaire score (Figure 8). The rapidly decreasing distribution allows removing a large number of less important attributes. A more abundant set of test subjects will allow us to continue the experiments.

Acknowledgements

We acknowledge the fruitful interaction with Dr. Michele Tagliati, Professor and Vice-Chairman in the Department of Neurology and Director of Movement Disorder at the Cedars-Sinai Medical Center for his assistance to the LIONSolver team in understanding the challenges, practical context, and open research issues related to Parkinson's disease. In addition we acknowledge the LION Lab at University of Trento (Italy) for providing the high-performance computational resources used for this study.

Bibliography

[1] R. Battiti. Using the mutual information for selecting features in supervised neural net learning. IEEE Transactions on Neural Networks, 5(4):537–550, 1994.

[2] R. Battiti and A. M. Colla. Democracy in neural nets: Voting schemes for accuracy. Neural Networks, 7(4):691–707, 1994.

[3] Roberto Battiti, Mauro Brunato, and Franco Mascia. Reactive Search and Intelligent Optimization. Operations research/Computer Science Interfaces. Springer Verlag, 2008. ISBN: 978-0-387-09623-0.

[4] Bernhard Schölkopf, Alex J Smola, Robert C Williamson, and Peter L Bartlett. New support vector algorithms. Neural computation, 12(5):1207–1245, 2000.

[5] V. N. Vapnik. The Nature of Statistical Learning Theory. Springer Verlag, 1995.

[6] LIONsolver, Inc. website: http://lionsolver.com/

Appendix

The following appendices contain more details about the techniques described in the previous pages.

A - Data cleaning and preprocessing

Because data are collected in a passive way the movement-related symptoms (shaking, rigidity, slowness of movement) can be shaded by other effects acting as "noise" sources (imagine a patient being in a moving car), which need to be filtered out as much as possible to make the "signal" emerge. To clean the data one could proceed by hand, trying to identify all possible external noise sources. We follow instead an unsupervised learning (clustering) approach: in addition to being effective the clustering approach has the advantages of automation, speed, flexibility to adapt to different conditions, robustness.

The data collected during the trial is analyzed and preprocessed in order to identify relevant data segments. Compass and GPS readings are used to isolate time windows in which the phone wasn't being moved or rotated in an excessive manner. Accelerometer readings are then averaged within these time windows (Section A.1). Further analysis shows the existence of local correlations between the data in Parkinson Disease (PD) subjects that do not appear in the Control Group (CG) subjects (Section A.2).

A.1 Clustering for Data Cleaning

Our experiments show that the most valuable data collected by the phone is the accelerometer reading, which can quantify the patient's tremors. The data provided has information about the Power Spectral Density (PDS) at various frequencies along each axis averaged over each second of sampling, on different frequencies values (continuous component, 1Hz, 3Hz, 6Hz and 10Hz); this information can be used to isolate traces of the involuntary tremors associated with Parkinson's disease.

The data collected by the cellphone is bound to be affected by external noise: for example, walking or sitting on a moving vehicle can influence the tremors recorded by the accelerometers. Therefore, the accelerometer data are filtered by considering other data recorded by the phone. Our approach to data cleaning relies on the compass and GPS records. The objective is to isolate the time windows in which the phone didn't change location and the subject was still. The importance of isolating such moments is underlined by the fact that tremors in Parkinson patients tend to be stronger when the subject is at rest.

The technique adopted to clean the dataset is to cluster the data with respect to the compass readings (pitch, roll and azimuth, describing the inclination of the device), the GPS data (longitude and latitude) and recording time. We perform this operation on each 1-hour recording session, isolating dense clusters and discarding the outliers. In this way we ensure that the relevant data are collected while the phone was lying in the same position and the subject was not moving.

Each recording session (at most one hour), has been analyzed by means of the K-means algorithm applied to the following features:

- the recording time;
- the values of the compass data (mean, absolute deviation, standard deviation and max deviation for the three components of the compass data: azimuth, pitch and roll);
- when available, the GPS location (defined by latitude, longitude and altitude). In our experiments we chose to identify K = 10 clusters per recording session. The number of clusters determines the granularity of the final dataset, and the validity of a cluster is determined with respect to two thresholds, one on the number of samples it contains and one on its density. K-means uses the mean Euclidean distance metric, with each coordinate appropriately scaled to account for different units of measures and technical issues (like the fact that 360 degree equals 0 degree when measuring angles). A valid cluster satisfies the following conditions:
- It contains a minimum of 10 samples: smaller clusters are indistinguishable from noise, and standard deviation measures are not significant with too few data items.
- the mean distance from the centroid must not exceed a fixed threshold which determines the acceptable level of mobility. The best threshold is determined automatically by searching for the best cross-validation results.

This automated clustering isolates contiguous time windows in which the phone has been in the same position and the subject was not moving at high speed.

The moduli of the three-axis accelerometer readings are computed for each frequency band, and the mean and standard deviation are computed for the whole cluster. Each cluster is thus summarized by eight numbers:

$$R = (m_{1\text{Hz}}, e_{1\text{Hz}}, m_{3\text{Hz}}, e_{3\text{Hz}}, m_{6\text{Hz}}, e_{6\text{Hz}}, m_{10\text{Hz}}, e_{10\text{Hz}})$$

where m_f is the arithmetic mean of the accelerometer power density in the frequency band around f (with f = 1Hz, 3Hz, 6Hz, 10Hz), while e_f is the corresponding standard deviation. The continuous component of the power distribution spectrum is not considered to focus on tremors, and thus the Earth's gravity acceleration component g is filtered out.

Figure 1 at page 2 describes the procedure: compass and GPS readings are used to delimit time windows, and in each time window accelerometer data are summarized into a vector.

At the end of the clustering procedure, every test subject is thus represented by a set of vectors in eight dimensions, each vector representing a valid measurement cluster found in the subject's history.

A.2 Exploiting Local Correlations

The mean value and the variance of the norm of the Power Spectral Density three-dimensional vector, computed on the clusters identified during the pre-processing phase gives information on the presence of tremors during the time window represented by the cluster. The motivation for our proposal is that recordings corresponding to Parkinson's tremors (or movement difficulties) will occupy regions of the space that are different from the recordings where no tremors has happened. Recordings from a Parkinson's Disease patient will fall in these regions more often than the ones of a control group member. As an example, Figure 2 at page 2 plots The $m_{3\text{Hz}}$ coordinate of valid time window summaries versus the $m_{10\text{Hz}}$ coordinate for PD and control subjects.

The plot features areas in which the majority of points clearly belongs to PD patients. Looking at lower frequency bands, on the other hand, the differences between the two classes are not so emphasized: this is in agreement with the fact that PD tremors tend to have a specific frequency in the 4Hz–6Hz range. For comparison, Figure 9 plots the mean PSD values at 3Hz and 1Hz: in this case the point clouds of the PD and control groups overlap almost completely and indicate a strong correlation between the two frequency ranges.

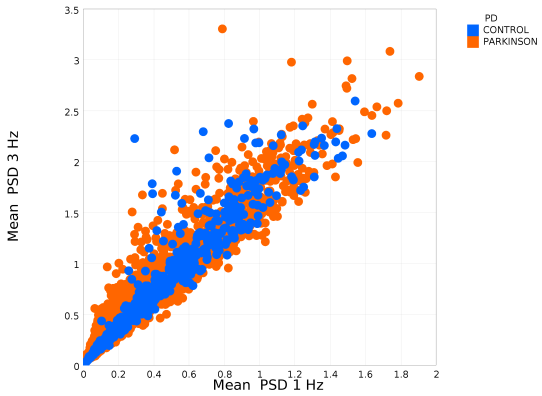

Figure 9 - Power spectral density at 1Hz and 3Hz at each valid time window for Parkinson and Control subjects: the two point clouds are much more overlapped than in other frequency ranges (see Figure 2 at page 2).

The analysis of a test subject's recordings shows that many points fall in an area of the space which is shared between the two classes of observations. Differently from the control group, the real PD Patients will have a fraction of recordings which fall in different sections of the space, occupied only by datapoints representing time intervals interested by tremors.

A simpler approach, such as describing a test subject using the mean of the PSD values at different frequencies, doesn't permit to separate the two classes with high confidence and leads to misclassifications, especially for the patients affected by lighter forms of the disease.

B Machine Learning for Robust and Scalable Identification

Input data for the machine learning algorithm are derived from the unsupervised learning phase described in Appendix A.1. The algorithm and the chosen experimental methodology are described and motivated in Section A.3.

B.1 Feature extraction by space quantization

As explained in Section A.1, every test subject is associated to a cloud of points in an 8-dimensional space. These clouds contain information about accelerometric powers at different frequencies and in different times. A measure of the overlap of different subjects' clouds is thus a convenient input to a supervised machine learning algorithm.

Not all dimensions are equally relevant; for instance, Figure 9 displays a clear correlation between the $m_{1\text{Hz}}$ and $m_{3\text{Hz}}$ coordinates. Let us consider, in general, a subset of $d = 1, \ldots, 8$ dimensions. To extract a set of comparable features for each test subject, each of the d selected coordinates is discretized into a small set of Q quantiles. For instance, if $Q = 4$, every coordinate is divided into 4 intervals such that every interval contains the same number of points. The quantile discretization is performed independently for each coordinate and taking into account all points of all test subjects. The diagram in Figure 10 exemplifies the discretization and

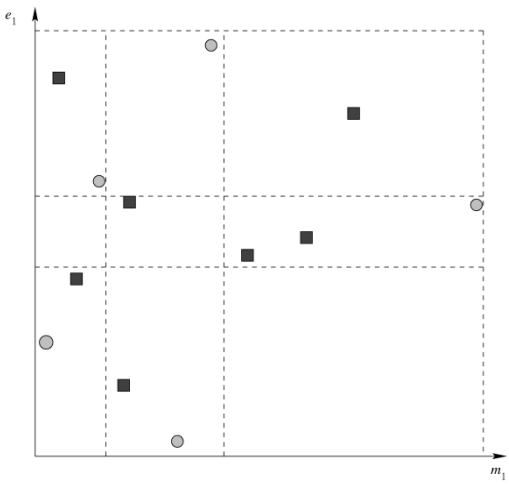

Figure 10 - Coordinate quantization. Each coordinate is divided into the same number of intervals so that every interval contains the same overall number of elements.

feature extraction procedure for two test subjects (whose points are represented by dark squares and light circles), two dimensions (e.g., $m_{1\mathrm{Hz}}$ and $e_{1\mathrm{Hz}}$) in place of eight, and $Q = 3$. In this simplified example, the interval division is such that each of the three horizontal intervals subtends the same number of points (four), and the same is true for the vertical axis intervals. Finally, the set of features of each test subject is obtained by counting the number of points in each of the 9 rectangular zones. Listing the intervals from left to right, and from bottom to top, the feature set of the first user (dark squares) is

(1,1,0,0,1,2,1,0,1),

while the feature set of the second user (light circles) is

(1,1,0,0,0,1,1,1,0).

The two feature sets are directly comparable and their normalized Euclidean distance is a good representation of how much the two point distributions overlap (occupy the same regions of the plane).

In the actual case, the *d*-dimensional space is divided into Qd regions. For large *Q* and *d*, given the possibly large number of regions, most of them will be empty. In order to accelerate the learning process and have smaller files, regions with no points whatsoever are not listed in the feature vector. For instance, considering the 16 test subjects, $d = 8$, and $Q = 4$, only 2106 of the $4^8 = 16384$ regions contain at least one point. Thus, the feature vectors have a fixed size of 2106 elements.

B.2 Support Vector Machines for small datasets and sparse feature vectors

The Support Vector Machine algorithm [5] typically performs well even for small datasets and sparse feature vectors. Given the small number of samples, we chose to give priority to the number of support vectors with respect to margin maximization. This can be achieved by the v-support vector classification method [4], that allows the tuning of parameter v as a lower bound to the ratio of support vectors to samples. The algorithm has been tested with a Radial Basis Function kernel:

$$K(x_i, x_j) = e^{-\gamma \|x_i - x_j\|}.$$

The setup was tested with the LIONsolver software [6].

B.3 Leave-one-out testing

Because of the small number of samples, the leave-one-out methodology is a robust way to use all apart one example for training while still obtaining a significant estimation of the generalization performance (the performance on new patients, not considered during the training phase). Each sample, in turn, is treated as a test instance, while all the others are used to train the system. In detail, the procedure is the following:

1. Set the number of successes to zero.
2. For every test subject $s = 1,\ldots,16$:
    1. Put all test subjects, with the exclusion of *s*, in the training set.
    2. Compute the quantile intervals on the training set (this step is important to avoid any influence of *s* on the training phase: because of the small number of samples, quantile distribution is strongly dependent on the data associated to each sample).
    3. Compute the features of all points but *s* and create the training set.
    4. Train the v-SVM with a RBF kernel on the training set.
    5. Compute the features of test subject s using the same quantile intervals. To this purpose, consider the leftmost interval of each coordinate to extend to $-\infty$ and the rightmost one to extend up to $+\infty$.
    6. Predict the category of test subject *s*. If the predicted category matches the actual one, increase the number of successes.

C. Results and analysis

Tests have been performed for various combinations of SVM setups (C-SVM against v-SVM; polynomial against RBF kernel; normalized or unnormalized data). The best results were obtained with the v-SVM with v = .8 and an

RBF kernel; the number of successes for different combinations of the remaining parameters is shown in Figure 5 at page 3. The changing parameters are:

- d — the number of coordinates of the point cloud space. If d = 3, the three mean values for the frequency bands in the 3Hz, 6Hz and 10Hz are used, while the corresponding standard errors are discarded; results are also reported for $d = 1$ using only the mean power in the 3Hz range.
- Q — the discretization level, given as the number of quantiles each dimension is partitioned in; results are shown for $Q = 2$, i.e., a binary partition with the median as a discriminant, and $Q = 5$, corresponding to a finer partition of the vector space;
- γ — the parameter of the RBF kernel, defining how much the similarity between samples decreases with distance; it also balances the sparsity of the vector, therefore in our case it cannot be too small (as shown by the results).

From the results in Figure 4 at page 3, we see that interestingly good results can be obtained from just one coordinate, and for a reasonably wide range of values for γ (consider that for $\gamma \geq .1$ all tests are 100% correct on the whole 16-sample set). The next space size, $d = 3$, consistently provides higher prediction levels at the expense of a slightly higher number of features.

Of course, as more and more test subjects are added to the system (see Appendix D.1 on scalability) we can expect the optimal number of features to grow as well.

The number of data points in the test subject's cloud is critical to classification. In fact, the single most misclassified subject (every time that the number of correct classifications is less than 16) is IRIS (positive class), who has a very low number of valid measurement intervals.

D. Further developments

D.1 Scalability for large number of patients

The small number of test subjects limits the possibility to assess a prediction technique. Our results are very promising and compatible with identification of Parkinson's disease with close to 100% performance but of course additional patients are needed for a more precise validation. The advantages offered by the availability of a large sample are, among the others:

- Better classification of borderline cases, for example in case of weak symptoms.
- Beyond yes/no classification, a probability measure or a score system can be used to obtain more detailed indications about the severity of the disease and the disease progression in time.
- Feature selection: A larger dataset yields higher confidence in selecting the most relevant features for the Machine Learning algorithm, which in turn helps reducing the learning task size and, once the most relevant features are identified, facilitates human interpretation of the machine's decisions.
- From the point of view of drug development for Parkinson's disease, a massive scale study of many patients can be extremely valuable to assess in a data-driven manner the effect of cure protocols on patients, possibly aiming ad individualized cure protocols, adapted to patient characteristics, possibly including genetic data. Better effectiveness and faster drug developments can be expected.

On the other hand, a large dataset poses some problems in the preprocessing phase. The size of the raw data files grows with the number of test subjects, but also grows in time for every subject. However, all preprocessing tasks described in this paper are highly parallelizable:

- The clustering phase described in Appendix A.1 is always performed on 1-hour (or less) time segments of an individual, therefore the K-means algorithm is always applied to datasets whose size is independent from the number of subjects or the overall monitoring time. This means that clustering tasks can be distributed among different machines with very little communication overhead (only data distribution and collection).
- The feature extraction phase depends in principle on all data: quantile calculation requires sorting each coordinate of the whole dataset. However, as the number of data item grows, subsampling of data can be applied with increasing confidence, so that the computational effort is contained
- The SVM training phase depends on both the number of subjects and on the number of features. The latter can be considered a constant which will be experimentally fixed from time to time and tuned by setting the number of dimensions d and the number of quantiles *Q*. Training tasks with millions of items and thousands of features are a viable challenge in Machine Learning.

D.1.1 Big Data and Cloud-Based Machine Learning

A suitable task for parallel computing is the parameter optimization of the SVM machine learning procedure. In fact, the best parameter values are correlated to the size and distribution of the sample data, and also to the number of extracted features.

However, the dependency of parameters from the task size is stochastic; therefore, an optimization procedure needs to run several training tasks with the same parameter set (e.g., by computing a k-fold cross validation or by randomly sampling the training set) before being able to assess a fitness value.

The training of a machine learning algorithm is a CPU-consuming task; concurrent training on multiple computers is possible with nearly linear speedups because the communication overhead required to distribute problem data and collect results becomes more and more negligible as the sample set sizes increase.

Preliminary experiments on 300 computers hosted on Amazon's EC^2 platform and based on the Message-Passing Interface (MPI) parallelization library show that speedup mainly depends on the slowest training time among all instances, therefore a near-linear parallelization speedup can be achieved by setting adequate timeouts that stop slowly progressing instances and assign them a low score to discourage the corresponding parameter values.

E. Statistical analysis

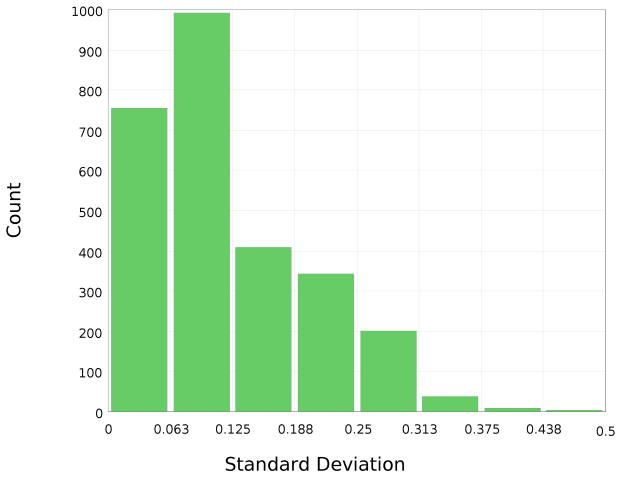

Figure 11 - Distribution of the standard deviation of accelerometer readings for subject APPLE

The Standard deviation from the mean value of the accelerometer is a good indicator of the presence of tremors in the data collected. The approach followed is to model the distribution of the norms of the three-dimensional standard deviation vectors, fitting them to a Gamma distribution. Figure 11 shows the distribution of the Standard Deviation values for the test subject APPLE.

The Standard deviation values found in each cluster identified in the data cleaning pre-processing phase are fitted to the Gamma distribution. Therefore, for each valid time frame a 3-dimensional tuple is obtained, containing the parameters of probability density function:

$$t = (\text{shape}, \text{location}, \text{scale}).$$

From the analysis of the distribution of the Gamma parameters, a strong correlation with the status of the test subjects visually appears. Figure 12 plots the shape parameter against the location parameter. It clearly shows a noticeable difference in the distribution between Control Group and Parkinson's Patients.

The interval considered for this analysis contains the majority of the observations, treating the remaining ones as outliers. The barcharts in Figure 13 and Figure 14 show how the location and shape parameters are distributed.

The mean value of the parameters of the Gamma distribution permits to represent a test subject in a three-dimensional space as the following tuple:

$$s = (\text{mean}_{\text{shape}}, \text{mean}_{\text{locatiob}}, \text{mean}_{\text{scale}}).$$

It is therefore possible to identify regions in the parameter space that are mainly occupied by Parkinson or Control data.

The parameters of the Probability density function characterizing each cluster are valuable features in recognizing the presence of tremors during a recording session. In fact a preliminary analysis on these features, through the representation of each test subject using the means of the parameters of the distributions corresponding to his recordings, produces the similarity map in Figure 6 at page 3.

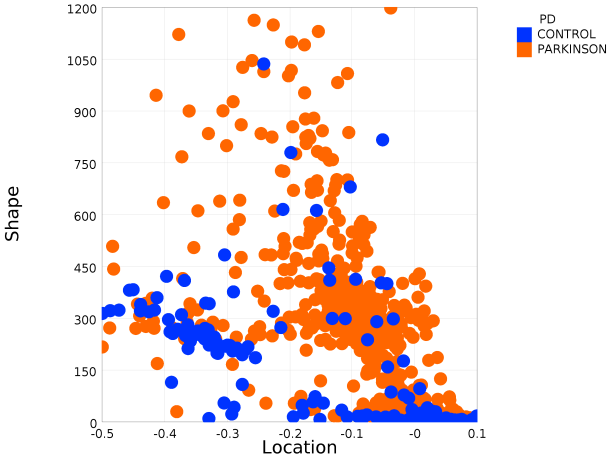

Figure 12 - Shape parameter vs. location parameter in the Gamma distribution fit of the standard deviation values

The figure, which maps similar data items in close positions trying to reproduce the high.dimensional data topology in two dimensions, shows how the two groups are clearly distinct, supporting the evidence of how the data collected passively can be used proficiently to distinguish the Parkinson's Disease tremors.

The test subject APPLE is very near to the group of real patients, this can be explained due to his age, which can possibly cause tremors similar to the ones caused by Parkinson disease. Subject CHERRY, on the other hand, is a PD subject, but her form of disease seems to be in a lighter stage, considering the information in the questionnaire.

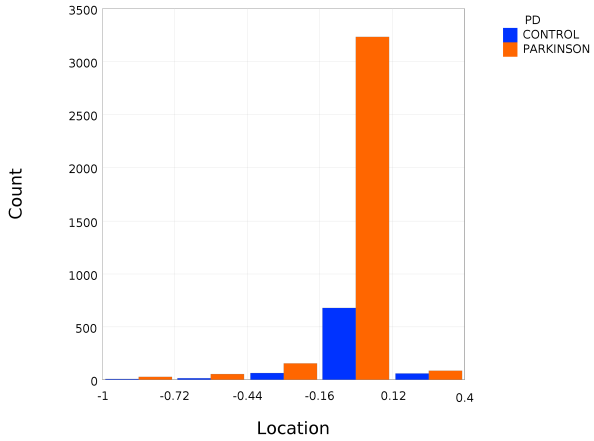

Figure 13 - Distribution of the Location parameter of the Gamma fit on standard deviation values.

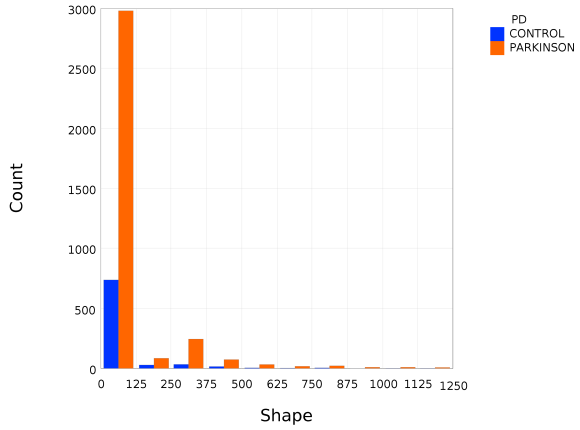

Figure 14 - Distribution of the shape parameter of the Gamma fit on the standard deviation values.