View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by AIS Electronic Library (AISeL)

Proceedings of the 51^{st} Hawaii International Conference on System Sciences | 2018

Big Data and Parkinson's Disease: Exploration, Analyses, and Data Challenges.

Mahalakshmi SenthilarumugamVeilukandammal
Iowa State University
maha10@iastate.edu

Dr.Sree Nilakanta
Iowa State University
nilakant@iastate.edu

Dr.Baskar Ganapathysubramanian
Iowa State University
baskarg@iastate.edu

Dr.Vellareddy Anantharam
Iowa State University
anantram@iastate.edu

Dr.Anumantha Kanthasamy
Iowa State University
akanthas@iastate.edu

Dr. Auriel A Willette
Iowa State University
awillett@iastate.edu

Abstract

In healthcare, a tremendous amount of clinical and laboratory tests, imaging, prescription and medication data are being collected. Big data analytics on these data aim at early detection of disease which will help in developing preventive measures and in improving patient care. Parkinson disease is the second-most common neurodegenerative disorder in the United States. To find a cure for Parkinson's disease biological, clinical and behavioral data of different cohorts are collected, managed and propagated through Parkinson's Progression Markers Initiative (PPMI). Applying big data technology to this data will lead to the identification of the potential biomarkers of Parkinson's disease. Data collected in human clinical studies is imbalanced, heterogeneous, incongruent and sparse. This study focuses on the ways to overcome the challenges offered by PPMI data which is wide and gappy. This work leverages the initial discoveries made through descriptive studies of various attributes. The exploration of data led to identifying the significant attributes. We are further working to build a software suite that enables end to end analysis of Parkinson's data (from cleaning and curating data, to imputation, to dimensionality reduction, to multivariate correlation and finally to identify potential biomarkers).

1.Introduction

Parkinson's Disease (PD) is a neurodegenerative disorder and millions of people suffer with it all over the world. The incidence of PD increases with the age growth, about 6.3 million people live with this disease. Especially, in developed country, the number of patients with PD has increased significantly in recent years. However, there are no methods which can measure the PD progression efficiently and accurately in its early stages [1]. The last known drug for Parkinson's disease was found in 1967.

Common symptoms in PD are muscular rigidity (inflexibility of muscles), shivering (vibration in upper and lower limbs or jaws), speech problem, expressionless face, Bradykinesia (slow movements), lethargy, postural instability (depression and emotional changes), involuntary movements, dementia (loss of memory), thinking inability and sleeping disorders. Various stages of Parkinson's disease are,

- Primary - Due to unknown reasons
- Secondary - Dopamine deficiency
- Hereditary- Genetic origin
- Multiple system atrophy - Degeneration of parts other than midbrain

For traditional PD assessment, Movement Disorder Society-sponsored Unified Parkinson Disease Rating Scale (MDS-UPDRS) is wildly used. To better understand PD progression and to identify potential biomarkers PPMI (Parkinson's Progression Markers Initiative) was set up. Clinical sites in the United States, Europe, Israel and Australia contribute to the comprehensive study. PPMI is funded by the Micheal J. Fox Foundation. PPMI collects clinical, biological and imaging data from multiple sites and

URI: http://hdl.handle.net/10125/50240
ISBN: 978-0-9981331-1-9
(CC BY-NC-ND 4.0)

disseminates it. This data is used to diagnose, track and predict PD and its progression.

Parkinson's data possess all the characteristics of big data, which are characterized by volume, variety, velocity, veracity, and value. From the context of Parkinson's data, these five Vs are further detailed as below.

- Volume – With more and more attributes being collected for the Parkinson's research and with the increase in participation of different cohorts through various initiatives, the volume of the data is growing.
- Variety –Parkinson's disease contains structured, text, images, audio and semi-structured data collected from the various smart fitness tracking devices
- Velocity-Velocity is depicted by the speed in which data is created, stored and processed. Nowadays real-time processing systems aid in real-time decision making.
- Veracity- Veracity deals with integrity of data. Data quality issues and reliability of the information are the key elements in veracity. Parkinson's data is heterogeneous, multi-source, incomplete, incongruent and sparse.
- Value- Extracting value from the data is the goal of big data analytics.

The goal of working with the Parkinson's data from the public databases is to find potential biomarkers thereby finding a cure for the disease.

Cleaning and curating data, however, to discover patterns from it is very challenging [2].

The main contribution of this study is to identify significant attributes that lead to PD. We first grouped 1358 unique attributes to six major categories. The data is then cleaned and curated. Redundancy in attributes was removed. Out of 2600 attributes, only 1358 were unique attributes. Descriptive studies of all these attributes were done. We wish to answer several questions in our research

- How can we discover the potential biomarkers of Parkinson's disease by using big data methodologies?
- Is it possible to use various machine learning algorithms to help in early detection of Parkinson's disease?
- What are the data that needs to be analyzed to discover the biomarkers of Parkinson's disease?
- How can we develop an interactive visualization that helps physicians understand the relations between various attributes that are a potential cause of Parkinson's disease?
- Is it feasible to scale the visualization for many user inputs? Does it yield the same result as the initial visualization with the training set?

In this study, we first tried understanding the attributes and created a metadata of the attributes. We did descriptive studies and computed the average of PD and HC for all the attributes to identify significant or important attributes. Curating the incomplete, heterogeneous data has proven to be the biggest challenge.

2. Related Work

In recent years, big data technologies are widely used in healthcare for earlier diagnosis of diseases and to provide better patient care. Dinov et al [3] illustrated bigdata's challenges and the role of big data technology in the biomedical field. They explore how the volume, variety, and velocity of biomedical data have tremendously increased. The challenges posed by biomedical data analysis is overcome by the pipeline environment. The pipeline is a crowd-based distributed solution for consistent management of these heterogeneous resources. The pipeline allows multiple (local) clients and (remote) servers to connect, exchange protocols, control the execution, monitor the states of different tools or hardware, and share complete protocols as portable XML workflows. As stated in their paper, *Laboratory of Neuro Imaging* (LONI) is one such pipeline environment for Parkinson's big data research. LONI seeks to improve understanding of the brain in health and disease.

Big Data analytics is applied on data collected by LONI from different sources. Machine learning techniques help to predict the PD at an earlier stage. Chen et al. [4] present an effective and efficient diagnosis system using fuzzy k-nearest neighbor (FKNN) for Parkinson's disease (PD) diagnosis. The proposed FKNN-based system is compared with the support vector machines (SVM) based approaches. To further improve the diagnosis accuracy for detection of PD, principle component analysis was employed. The effectiveness of the proposed system has been rigorously estimated on a PD dataset in terms of classification accuracy, sensitivity, specificity and the area under the receiver operating characteristic (ROC) curve. Experimental results have demonstrated that the FKNN-based system greatly outperforms SVM-based approaches and other methods in the literature. Gracy et al. [5] have discussed the four types of classifiers namely, Naive Bayes, Random tree, J48 and decision tree. Shivering hands, legs, arms or jaws and emotional changes are the factors considered in the study.

In the era of big data, the data quality is a big challenge when applying machine learning techniques and derive value from it. Ramentol et.al. [6] have stated that imbalanced data is a common problem in classification. Their paper proposes a new hybrid method for preprocessing imbalanced data-sets through the construction of new samples, using the Synthetic Minority Oversampling Technique together with the application of an editing technique based on the Rough Set Theory and the lower approximation of a subset. The proposed method has been validated by an experimental study showing good results using C4.5 as the learning algorithm. Cho et al. [7] proposed a system for combining principal component analysis (PCA) with linear discriminant analysis (LDA). They proposed a gait analysis system which can detect the gait pattern of Parkinson's disease using computer vision. Dinov et al. [8] introduces methods for rebalancing imbalanced cohorts and utilizes a wide spectrum of classification methods to generate consistent and powerful phenotypic predictions. It generates reproducible machine learning based classification that enables the reporting of model parameters and diagnostic forecasting based on new data.

Data collected in clinical studies is complex. Data visualization is paramount to enhance the understanding of data. Maciejewski et al. [9] have provided visual analytics systems to users to explore trends in their data. Linked views and interactive displays provide insight into correlations among people, events, and places in space and time. Furthermore, this study helps facilitate forecasting, as it has created a predictive visual analytics toolkit that provides researchers with linked spatiotemporal and statistical analytic views. Though there are several machine learning algorithms that have been implemented on different datasets on Parkinson there is no software developed to visually explore the correlations among various attributes and PD progression. Our study aims to visualize the risk factors and their relationship to PD.

3.Research Goals

Data collected from PPMI study consists of clinical, biological and imaging data of various patients. There are 2600 attributes and the number is constantly increasing as it is an ongoing study. This paper addresses the general challenges of data curation, munging, aggregation, and preliminary descriptive analyses of the PPMI data. This paper provides the results of the preliminary analyses.

Cleaning and curating the data is the biggest challenge. Each file was taken individually, redundant and administrative data that was not required for the study was removed. Aggregating these wide attributes together creates a huge sparse matrix. Finding the correlation between various attributes and visualizing them is the goal of this paper. Merits – This framework with a simple interactive visualization will abstract people from sophisticated mathematics to provide a simplified and understandable version of the disease to the life style of a common man.

After doing the initial analysis we have formulated two long term objectives for our study. Long term objective 1 focuses on tools for curating and imputing missing data using a set of novel algorithms.

Long term objective 2 consists of tools for reducing the dimensionality of the post-imputation data. End-users may effortlessly deploy several dimensionality reduction strategies, visually explore, and pick the most insightful approach.

4. Data

The PPMI study dataset is disseminated by PPMI Bioinformatics Core at the University of Southern California. This database includes clinical, biological and imaging data collected at various participating sites. PPMI also collects biologic specimens including urine, plasma, serum, cerebrospinal fluid, DNA, and RNA. The complete PPMI data set includes Biospecimen (ex: Lab reports, Blood sample), Imaging, Medical History, Subject Characteristics (ex: Demographics), Motor Assessment, and Non-Motor Assessment.

Table 1. Details of Files in PPMI

No of Files (CSV/Tables)	92
No of Files containing Administrative Data	12
No of Files containing Clinical, Questionnaire data	80

Table 2. Details of Various Attributes in PPMI

Total Number of Attributes	2600
Total Number of Unique Attributes	1358

Numerical Attributes	779
Categorical Attributes	1316
Time	47
Date	458

The dataset obtained from PPMI for our study consists of 1479 patients. PD are patients with Parkinson’s disease and 418 PD patients were considered in our study. Healthy control (HC) was 172 in number. We considered 418 PD, 172 HC and 62 Prodromal patients (In medicine, a prodrome is an early sign or symptom or set of signs and symptoms, which often indicate the onset of a disease before more diagnostically specific signs and symptoms develop). These totals up to 652 out of 1479 patients. The remaining 827 are genetic cohorts and genetic registry patients will be considered in the future research. Genetic Cohort PD, Genetic Cohort Unaffected, Genetic Registry PD, and Genetic Registry Unaffected are the other Cohorts in the dataset, but these cohorts were not included in the current study.

Table 3. Details of Patient Status

HC - Healthy Control	172
PD - Parkinson's Disease	418

4.1. Data Categorization

Data from various files are categorized into six major categories such as

a) Biospecimen (ex: Lab reports, Blood sample),

b) Imaging (ex: DaTscan imaging, Magnetic Resonance Imaging)

c) Medical History (ex: General medical history, General neurological exam, General physical exam, Pregnancy forms, Neurological exam cranial nerves)

d) Subject Characteristics (ex: demographics, PPMI took place at clinical sites in the United States, Europe, Israel, and Australia),

e) Motor Assessment (ex: assessment of tremor with bradykinesia, assessment of tremors in tongue, jaw, lower lip, hand or in the leg/foot. Movement Disorder Society (MDS) offers Unified Parkinson’s Disease Rating Scale (UPDRS) which guides in the motor assessment), and

f) Non-Motor Assessment (ex: assessment of verbal learning, semantic fluency and sleepiness scale are some of the non-motor assessment tests).

Figure 1 is visual representation of the file and its category. This categorization helps us understand the attributes better on a high level.

Table 4. Details of Various Data Categories after Data was Analyzed and Cleaned

Category	Number of Files
Biospecimen	11
Imaging	9
Medical History	14
Motor Assessment	11
Non-Motor Assessment	16
Subject Characteristics	5

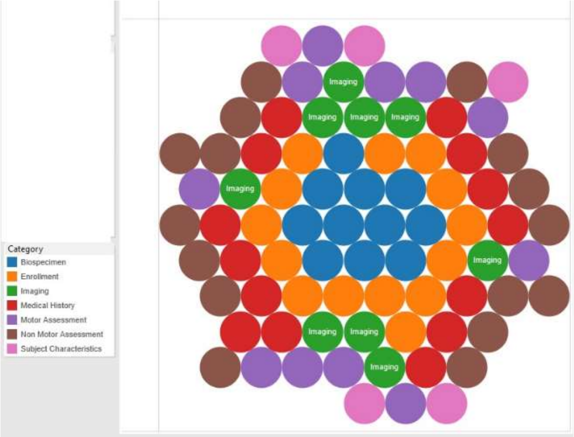

Figure 1. Data Categories

5. Analyses and Findings

The individual data files were cleaned and redundant data were removed. Out of initial 92 files(as mentioned in table 1) , only 66 files (as mentioned in table 4) contained the features associated with Parkinson’s disease. The administrative data about the enrollment status of different cohorts were excluded. The final list after removing the redundant and administrative attributes had 978 attributes. A descriptive study of all attributes was done. Mean, median, minimum, maximum, mode and standard deviation of all the

978 attributes were calculated. After the preliminary data exploration, we studied the correlation of different attributes with the patient status.

Data was mapped to the same standardized names. (ex: Some files had attributes called subject id and some had attributes called Patient Number). Once the data was standardized. All the data was loaded into PostgreSQL database. The data model of the database has the six main categories such as biospecimen, imaging, motor assessment, non-motor assessment, medical history and subject characteristics.

Metadata file containing the information about the data was created.

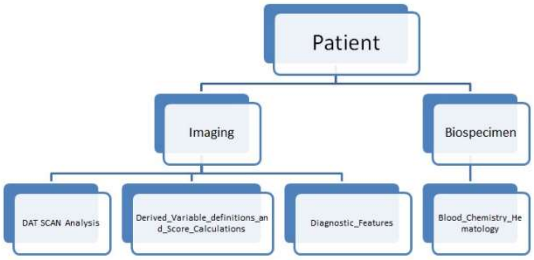
Patient
Imaging
Biospecimen
DAT SCAN Analysis
Derived_Variable_definitions_and_Score_Calculations
Diagnostic_Features
Blood_Chemistry_Hematology

Figure 2. A High-Level Overview of the Data Model

All the data from PPMI after cleaning was loaded into a PostgreSQL database. Figure 2 gives a high level data model of the database. This indicates how the tables were created and data was loaded.

5.1. Significant Attributes

Once all the attributes are aggregated, the average value for Parkinson's disease PD and healthy control HC cohort was calculated. The difference in the value was normalized. Figure 3 illustrates how the average value of PD and HC appear after normalization. The values are between 0 and 1 and easy to compare. The difference in value was visualized in Tableau which was connected to the database as illustrated in figure 4. The attributes with a significant difference in value were identified. In a high dimensional dataset discovering the important features is crucial. The study results demonstrated the attributes in MDS_UPDRS_Part_III had the significant attributes.

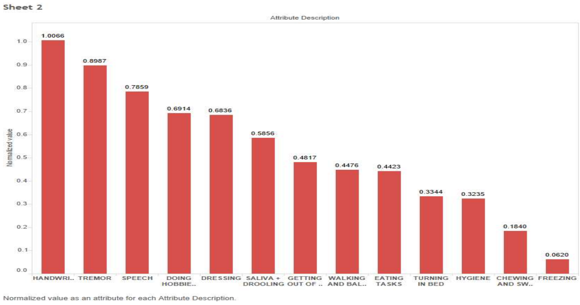

Figure 3. Normalized Data

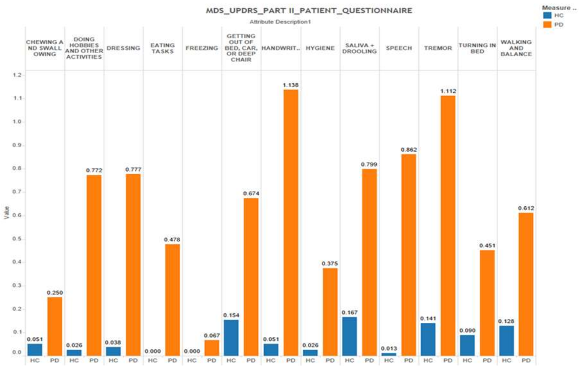

Figure 4. Attributes with Significant Difference between PD and HC Values.

We also, explored the correlations among the attributes using circos visualizations. See figure 5. Circos visualization is an interactive tool in that we can isolate the relationship between one attribute and all other attributes, displaying the strength of the correlation as a measure of the width of the flare. An interactive version will be provided later.

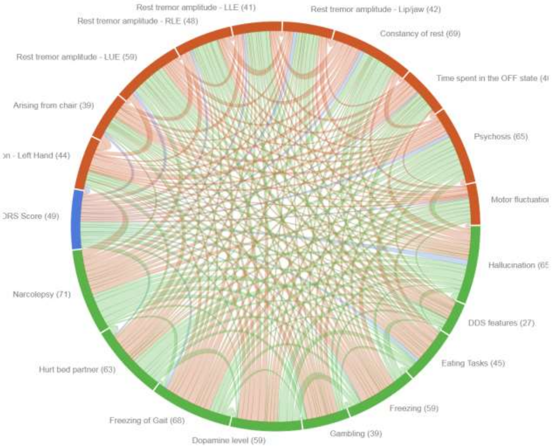

Figure 5. Circos Visualization with Significant Attributes.

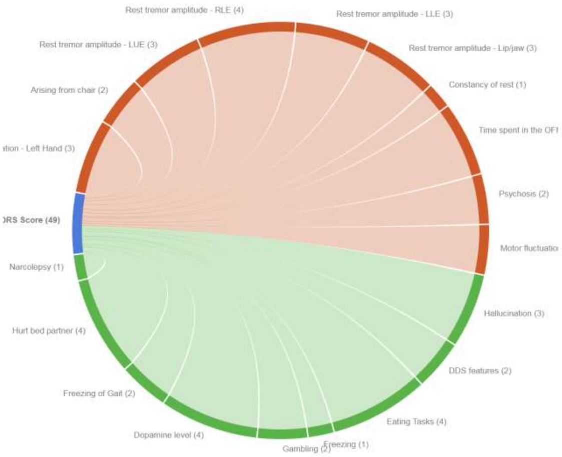

Figure 6. Circos Visualization Displaying Correlation between Attributes.

Figure 6. illustrates how UPDRS (Unified Parkinson's Disease Rating Scale) is highly correlated with bradykinesia, postural instability, hallucination and speech. Furthermore, the graph is interactive as we can hover-over the visualization to find the correlations of any other attributes with others.

6. Challenges in Dataset

The dataset from PPMI is wide with 2600 attributes. The complexity is furthermore increased as it is a time series data. Curating and stacking the data is a big challenge. The data is incomplete, imbalanced and incompatible. When data is aggregated together there are lots of missing value.

Only 30% of data is available. Imputing the missing value is of paramount importance for implementing various machine learning algorithms on the data to identify the potential biomarkers.

To overcome the challenges posed by human clinical datasets we are researching on a set of novel algorithms. (a) Singular Value Decomposition type imputation, (b) gappy Tensor decomposition, and (c) standard knn based imputation. The latter approach, prevalent in biomedical research, will serve as a benchmark. The two former methods (especially the Tensor decomposition) have shown phenomenal ability to impute complex datasets in engineering problems and will translate into novel approaches for biomedical data.

7. Conclusion

Currently, data from clinical and behavioral studies of PD are growing rapidly and with little knowledge or coordination of attributes collected. Understanding the importance of each attribute collection to PD detection and treatment is important and our work helps highlight the challenges in data quality and tools to improve the same. Future work can be extended by allowing researchers to add additional attributes and determine their role in PD.

8. Reference:

[1] S.L.Kowal, T.M. Dall, R. Chakrabarti, et al.," The current and projected economic burden of Parkinson's disease in the United States.", Movement Disorders. 2013 Mar;28(3):311-8.

[2] M. Minelli, M. Chambers, A. Dhiraj, Big Data Big Analytics: Emerging business intelligence and analytics trend for today's businesses, Somerset, US: Wiley, Feb 2013.

[3] I.D. Dinov, P. Petrosyan, Z. Liu, et al.," The perfect neuroimaging-genetics- computation storm: the collision of petabytes of data, millions of hardware devices and thousands of software tools.", Brain Imaging and Behavior, 2014;8(2):311–22.

[4] H.L.Chen, C.C. Huang, X.G.Yu, et al. "An efficient diagnosis system for detection of Parkinson's disease using fuzzy k-nearest neighbor approach.", Expert Systems with Applications 40, (2013) 263–271.

[5] S. Hariganesh and S. Gracyannamary." Comparative study of Data Mining Approaches for Parkinson's Disease.", IJARCET september 2014, Vol: 3, Issue 9.3062-8

[6] E. Ramentol, Y.Caballero, R.Bello,et al., " SMOTE-RSB*: a hybrid preprocessing approach based on oversampling and under-sampling for high imbalanced data -sets using SMOTE and rough sets theory.", Knowledge and Inf ormation Systems. 2012;33(2):245–65.

[7] T.R. Morris, C.Cho, V. Dilda, et al. "Clinical assessment of freezing of gait in Parkinson's disease from computer-generated animation.", gait posture, 2013 Jun;38(2):326-9.

[8] I.D. Dinov, B. Heavner, M. Tang, et al. "Predictive Big Data Analytics: A Study of Parkinson's Disease Using Large, Complex, Heterogeneous, Incongruent, Multi-Source and Incomplete Observations.", PLOS (2016).

[9] R. Maciejewski, R.Hafen ,S.Rudolph , et al." Forecasting hotspots -A predictive analytics approach. Visualization and Computer Graphics", IEEE Transactions on. 2011;17(4):440–53.