DETECTION OF PARKINSON DISEASE BY USING BEST ACCURACY IN MACHINE LEARNING

S. KANMANI¹, G. LAVANYA², D. ANAND JOSEPH DANIEL³

¹Student, BE Computer science and engineering, Anand Institute of Higher Technology, Chennai, India
²Student, BE Computer Science and Engineering, Anand Institute of Higher Technology, Chennai, India
³Professor, BE Computer Science and Engineering, Anand Institute of Higher Technology, Chennai, India

Abstract:

Parkinson’s disease is the most prevalent neurodegenerative disorder affecting more than 10 million people worldwide. There is no single test which can be administered for diagnosing Parkinson’s disease. Because of these difficulties, to investigate a machine learning approach to accurately diagnose Parkinson’s, using a given dataset. To prevent this problem in medical sectors have to predict the disease affected or not by finding accuracy calculation using machine learning techniques. The aim is to investigate machine learning based techniques for Parkinson disease by prediction results in best accuracy with finding classification report. The analysis of dataset by supervised machine learning technique (SMLT) to capture several information’s like, variable identification, uni-variate analysis, bi-variate and multi-variate analysis, missing value treatments and analyze the data validation, data cleaning/preparing and data visualization will be done on the entire given dataset. To propose, a machine learning-based method to accurately predict the disease by speech and tremor symptoms by prediction results in the form of best accuracy from comparing supervise classification machine learning algorithms. Additionally, to compare and discuss the performance of various machine learning algorithms from the given transport traffic department dataset with evaluation classification report, identify the result shows that the effectiveness of the proposed machine learning algorithm technique can be compared with best accuracy with precision, Recall and F1 Score.

Keywords: dataset, Machine learning-Classification method, python, Prediction of Accuracy result.

1. INTRODUCTION

Machine learning is to predict the future from past data. Machine learning (ML) is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of Computer Programs that can change when exposed to new data and the basics of Machine Learning, implementation of a simple machine learning algorithm using python. Process of training and prediction involves use of specialized algorithms. It feed the training data to an algorithm, and the algorithm uses this training data to give predictions on a new test data. Machine learning can be roughly separated in to three categories. There are supervised learning, unsupervised learning and reinforcement learning. Supervised learning program is both given the input data and the corresponding labeling to learn data has to be labeled by a human being beforehand. Unsupervised learning is no labels. It provided to the learning algorithm. This algorithm has to figure out the clustering of the input data. Finally, Reinforcement learning dynamically interacts with its environment and it receives positive or negative feedback to improve its performance.

Data scientists use many different kinds of machine learning algorithms to discover patterns in python that lead to actionable insights. At a high level, these different algorithms can be classified into two groups based on the way they “learn” about data to make predictions: supervised and unsupervised learning. Classification is the process of predicting the class of given data points. Classes are sometimes called as targets/ labels or categories. Classification predictive modeling is the task of approximating a mapping function from input variables (X) to discrete output variables (y). In machine learning and statistics, classification is a supervised learning approach in which the computer program learns from the data input given to it and then uses this learning to classify new observation. This data set may simply be bi-class (like identifying whether the person is male or female or that the mail is spam or non-spam) or it may be multi-class too. Some examples of classification problems are: speech recognition, handwriting recognition, biometric identification, document classification etc. Supervised Machine Learning is the majority of practical machine learning uses supervised learning. Supervised learning is where have input variables (X) and an output variable (y) and use an algorithm to learn the mapping function from the input to the output is \( y = f(X) \). The goal is to approximate the mapping function so well that when you have new input data (X) that you can predict the output variables (y) for that data. Techniques of Supervised Machine Learning algorithms include logistic regression, multi-classes, DecisionTrees and support vector machines etc. Supervised learning requires that the data used to train the
algorithm is already labeled with correct answers. Supervised learning problems can be further grouped into Classification problems. This problem has as goal the construction of a succinct model that can predict the value of the dependent attribute from the attribute variables. The difference between the two tasks is the fact that the dependent attribute is numerical for categorical for classification. A classification model attempts to draw some conclusion from observed values. Given one or more inputs a classification model will try to predict the value of one or more outcomes. A classification problem is when the output variable is a category, such as “red” or “blue”.

LITERATURE SURVEY

LITERATURE SURVEY 1

CONCEPT USED
One to recognize the presence and severity of motor fluctuations in patients with Parkinson's disease. To identify motor patterns of primary and secondary movement disorders in Parkinson Disease(PD), such as tremor, rigidity, dyskinesia, akinesia and dystonia in a manner that is both objective and automatic.

{DATA MINING TECHNIQUES TO DETECT MOTOR FLUCTUATIONS IN PARKINSONS DISEASE}

LITERATURE SURVEY 2

CONCEPT USED
To use a weakly supervised machine learning framework for automatic detection of Parkinson’s Disease motor symptoms in daily living environments. Our primary goal is to develop a monitoring system capable of being used outside of controlled laboratory settings.

{DETECTING PARKINSONS' SYMPTOMS IN UNCONTROLLED HOME ENVIRONMENTS : A MULTIPLE INSTANCE LEARNING APPROACH}

LITERATURE SURVEY 3

CONCEPT USED
A more efficient system is required to cope with the growing world population and increased life expectancy, which is associated with a marked prevalence of chronic neurological disorders such as Parkinson’s disease. One possible approach to meet this demand is a laterally distributed platform such as the Internet of Things.

{AN EMERGING ERA IN THE MANAGEMENT PARKINSONDISEASE:WEARABLE TECHNOLOGY AND THE INTERNET OF THINGS}

LITERATURE SURVEY 4

CONCEPT USED
One of the most typical clinical hallmarks of PD is handwriting deterioration, usually the first manifestation of PD. The aim of this study is twofold: (a) to find a subset of handwriting features suitable for identifying subjects with PD and (b) to build a predictive model to efficiently diagnose PD.

{DECISION SUPPORT FRAMEWORK FOR PARKINSON DISEASE BASED ON NOVEL HANDWRITING MARKERS}

LITERATURE SURVEY 5

CONCEPT USED
Current clinical methods of Parkinson’s disease assessment are known to be subjective and potentially error prone. This paper presents a preliminary investigation into the development of a novel approach for the assessment of Parkinsonian tremor.

{ASSESSMENT AND VISUALIZATION OF PARKINSON DISEASE}
I. PREPARING DATASET

The dataset is now supplied to machine learning model on the basis of this data set the model is trained. Every new patient’s details filled at the time of application form acts as a test data set. After the operation of testing, model predict whether the new patients are a fit case for affected or not.

Gathering Data

Two set of data used:
A. Speech Dataset
B. Tremor Dataset

Tremor dataset
From all subjects, three types of handwriting recordings (Static Spiral Test (SST), Dynamic Spiral Test (DST) and Stability Test on Certain Point (STCP)) are taken. Also the drawings of spirals belongs to the PWP are included in the dataset.

Speech dataset
The Multi-Dimensional Voice Program (MDVP) is a well-established software program used for quantitative acoustic signal assessment of voice quality. The MDVP calculates a number of acoustic parameters including shimmer, short-term perturbations of the amplitude, and jitter, short-term perturbations of the frequency.

Unified Parkinson's Disease Rating Scale (UPDRS) report

II. EXISTING SYSTEM

Scan without evidence of dopaminergic deficit (SWEDD) are subjects clinically diagnosed as Parkinson’s disease (PD) patients although the SPECT imaging does not show any negro striatal abnormality. It defined five models of machine learning were used to carry out binary classification (healthy control/PD) using clinical assessment and image-derived features applied thereafter on SWEDD group as a potential application of motor and non-motor features in understanding Parkinson disease characteristic in this group. The nested cross-validation was an essential component to select reliable models. A high accuracy was achieved for the five models (75.4% - 78.4% for motor features and 71% - 82.2% for non-motor features) in binary classification (HC Vs PD). Accordingly, It demonstrate the suitability and usefulness of ML models to carry out binary classification of SPECT data. Cross all models applied on SWEDD group, 17.6% of patients were classified as PD motor disorder lookalikes, 27.4% were classified as having a beginning non-motor abnormality of PD and 3.9% were classified as having both motor and non-motor PD features. However, the interpretability of SWEDD predicted condition must be carefully considered.

To demonstrate the feasibility of evaluating machine learning algorithms to classify PD subjects from HC. The nested cross-validation used in this work was an essential component to select reliable models. The SVM and MLP models slightly outperform the other models using motor features and so was the K-NN using non-motor features. The retained models were applied on SWEDD subjects to predict their conditions as a PD patient with motor disorder or as a PD patient with non-motor abnormality. The fact that the ground-truth of this group is unavailable makes the evaluation the performance of these models on this group awkward. However, it suggests that a merit of investigation by using our models might be suitable for a follow-up.

The five models were applied on SWEED as a potential application of motor and non-motor features in understanding Parkinson disease characteristic in this group. SVM model was applied on the PC1 and PC2 scores and non-motor features and provided a classification of SWEDD as follow: a 57% of SWEDD patients were classified as PD motor disorder lookalikes by the SVM_M model whereas 51% were classified as having a beginning non-motor abnormality of PD by the SVM_NM. Importantly, 27% of SWEDD patients were classified as mimicking both PD motor and non-motor disorder and abnormality cross the two models. RF model was applied on the two subsets of features (motor and non-motor) and provided a classification of SWEDD as follow: a 43% of SWEDD patients were classified as PD with motor disorder by the RF_M model and 54% were classified as having a beginning non-motor abnormality of PD by the RF_NM. Cross the two models, 25% of SWEDD patients were classified as having both PD features. Similarly, K-NN model was applied on the two set of features and provided a classification of SWEDD as follow: a 25% of SWEDD patients were classified as PD motor disorder lookalikes by the K-NN_M model and 63% were classified as having a beginning non-motor abnormality of PD by the K-NN_NM model. Cross the two models, 24% of SWEDD patients were classified as having both PD features.

The MLP model provided a classification of SWEDD as follow: a 47% of SWEDD patients were classified as PD motor disorder lookalikes by the MLP_M model and 41% were classified as having a beginning non-motor abnormality of PD by MLP_NM model. Cross the two models, 26% of SWEDD patients were classified as having both PD features. The LR model provided a classification of as follow: a 47% of SWEDD patients were classified as PD motor disorder lookalikes by the first LR_M model and 38% were classified as having a beginning non-motor abnormality of PD by the LR_NM model. Importantly, 27% of SWEDD patients were classified as having both PD features.
abnormality of PD by the second model. Cross the two models, 24% of SWEDD patients were classified as having both PD features.

**Drawbacks:**

- There are known disadvantages associated with the use of sampling to implement cost-sensitive learning. The disadvantage with under sampling is that it discards potentially useful data. The main disadvantage with oversampling, from our perspective, is that by making exact copies of existing examples, it makes over fitting likely.
- In fact, with oversampling it is quite common for a learner to generate a classification rule to cover a single, replicated, example. A second disadvantage of oversampling is that it increases the number of training examples, thus increasing the learning time.

### III. PROPOSED SYSTEM

**Exploratory Data Analysis of Parkinson disease Prediction**

This analysis is not meant to be providing a final conclusion on the reasons leading to medical hospitals of detecting Parkinson as it doesn't involve using any inferential statistics techniques/machine learning algorithms. Machine learning supervised classification algorithms will be used to give the given dataset and extract patterns, which would help in predicting the likely affected or not, thereby helping the hospitals for making better decisions of detecting disease in the future. Multiple datasets from different sources would be combined to form a generalized dataset, and then different machine learning algorithms would be applied to extract patterns and to obtain results with maximum accuracy.

**Data Wrangling:**

In this section of the report, you will load in the data, check for cleanliness, and then trim and clean your dataset for analysis. Make sure that you document your steps carefully and justify your cleaning decisions.

**Training the Dataset**

- The first line imports iris data set which is already predefined in sklearn module. Iris data set is basically a table which contains information about various varieties of iris flowers.
- For example, to import any algorithm and train_test_split class from sklearn and numpy module for use in this program.
- Then we encapsulate load_data() method in data_dataset variable. Further we divide the dataset into training data and test data using train_test_split method. The X prefix in variable denotes the feature values and y prefix denotes target values.
- This method divides dataset into training and test data randomly in ratio of 67:33. Then we encapsulate any algorithm.
- In the next line, we fit our training data into this algorithm so that computer can get trained using this data. Now the training part is complete.

**Testing the Dataset**

- Now, it has dimensions of a new flower in a numpy array called ‘n’ and to predict the species of this flower. To do this using the predict method which takes this array as input and spits out predicted target value as output.

So the predicted target value comes out to be 0. Finally to find the test score which is the ratio of no. of predictions found correct and total predictions made. To do this using the score method which basically compares the actual values of the test set with the predicted values.

### IV General Properties

Create cells freely to explore the given data and it should not perform too many operations in each cell. One option that can take with this is to do a lot of explorations in an initial notebook. These don't have to be organized, but make sure you use enough comments to understand the purpose of each code cell. Then, after done with your analysis,
create a duplicate notebook where it will trim the excess and organize steps so that have a flowing, cohesive report and make sure that informed on the steps that are taking in your investigation. Follow every code cell, or every set of related code cells, with a markdown cell to describe to the reader what was found in the preceding cell. Try to make it so that the reader can then understand what they will be seeing in the following cell.

Advantages:
- To save doctors risk and increasing patient appointments.
- Easy to predicting the diagnose Parkinson with doctors can detecting the patient testing result time is reduced.

Application:
A hospital wants to automate the diagnosis of patient (real time) based on the patient detail provided while filling online/ offline application form. To automate this process, they have given a problem to identify the patient segments, those are have detecting disease to list to show doctors.

**ALGORITHM AND TECHNIQUES**

Algorithm Explanation
In machine learning and statistics, classification is a supervised learning approach in which the computer program learns from the data input given to it and then uses this learning to classify new observation. This data set may simply be bi-class or it may be multi-class too. Some examples of classification problems are: speech recognition, handwriting recognition, biometric identification, document classification etc. In Supervised Learning, algorithms learn from labeled data. After understanding the data, the algorithm determines which label should be given to new data based on pattern and associating the patterns to the unlabeled new data.

Used Python Packages:
- **sklearn:**
  - In python, sklearn is a machine learning package which include a lot of ML algorithms.
  - Here, we are using some of its modules like train_test_split, DecisionTreeClassifier or Logistic Regression and accuracy_score.
- **NumPy:**
  - It is a numeric python module which provides fast maths functions for calculations.
  - It is used to read data in numpy arrays and for manipulation purpose.
- **Pandas:**
  - Used to read and write different files.
  - Data manipulation can be done easily with data frames.
- **Matplotlib:**
  - Data visualization is a useful way to help with identify the patterns from given dataset.
  - Data manipulation can be done easily with data frames.

Data Visualization types:
- Introduction to Matplotlib
- Line Plot
- Bar Chart
- Histogram Plot
- Box and Whisker Plot
- Scatter Plot

**Bar Chart:**
A bar chart is generally used to present relative quantities for multiple categories. The x-axis represents the categories and are spaced evenly. The y-axis represents the quantity for each category and is drawn as a bar from the baseline to the appropriate level on the y-axis.
Box and Plot:

A box and whisker plot, or boxplot for short, is generally used to summarize the distribution of a data sample. The x-axis is used to represent the data sample, where multiple boxplots can be drawn side by side on the x-axis if desired. The boxplot is a graphical technique that displays the distribution of variables. It helps us see the location, skewness, spread, tile length and outlying points. The boxplot is a graphical representation of the Five Number Summary.

Heat map

A heat map is a graphical representation of data where the individual values contained in a matrix are represented as colors. It is a bit like looking a data table from above. It is really useful to display a general view of numerical data, not to extract specific data point. It is quite straight forward to make a heat map, as shown on the examples below.
Fig: Heat map diagram of given dataset for tremor

However be careful to understand the underlying mechanisms. You will probably need to normalize your matrix, choose a relevant color palette, use cluster analysis and thus permute the rows and the columns of the matrix to place similar values near each other according to the clustering.

Logistic Regression

It is a statistical method for analysing a data set in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). The goal of logistic regression is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables. Logistic regression is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.).

Decision Tree

It is one of the most powerful and popular algorithm. Decision-tree algorithm falls under the category of supervised learning algorithms. It works for both continuous as well as categorical output variables. Assumptions of Decision tree:

- At the beginning, we consider the whole training set as the root.
- Attributes are assumed to be categorical for information gain, attributes are assumed to be continuous.
- On the basis of attribute values records are distributed recursively.
- We use statistical methods for ordering attributes as root or internal node.

Decision tree builds classification or regression models in the form of a tree structure. It breaks down a data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. A decision node has two or more branches and a leaf node represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data. Decision tree builds classification or regression models in the form of a tree structure. It utilizes an if-
then rule set which is mutually exclusive and exhaustive for classification. The rules are learned sequentially using the training data one at a time. Each time a rule is learned, the tuples covered by the rules are removed.

This process is continued on the training set until meeting a termination condition. It is constructed in a top-down recursive divide-and-conquer manner. All the attributes should be categorical. Otherwise, they should be discretized in advance. Attributes in the top of the tree have more impact towards in the classification and they are identified using the information gain concept. A decision tree can be easily over-fitted generating too many branches and may reflect anomalies due to noise or outliers.

**K-Nearest Neighbor (KNN)**

K-Nearest Neighbor is a supervised machine learning algorithm which stores all instances correspond to training data points in n-dimensional space. When an unknown discrete data is received, it analyzes the closest k number of instances saved (nearest neighbors) and returns the most common class as the prediction and for real-valued data it returns the mean of k nearest neighbors. In the distance-weighted nearest neighbor algorithm, it weights the contribution of each of the k neighbors according to their distance using the following query giving greater weight to the closest neighbors.

Usually KNN is robust to noisy data since it is averaging the k-nearest neighbors. The k-nearest-neighbors algorithm is a classification algorithm, and it is supervised: it takes a bunch of labeled points and uses them to learn how to label other points. To label a new point, it looks at the labeled points closest to that new point (those are its nearest neighbors), and has those neighbors vote, so whichever label the most of the neighbors have is the label for the new point (the “k” is the number of neighbors it checks). Makes predictions about the validation set using the entire training set. KNN makes a prediction about a new instance by searching through the entire set to find the k “closest” instances. “Closeness” is determined using a proximity measurement (Euclidean) across all features.

**Random Forest**

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees’ habit of over fitting to their training set. Random forest is a type of supervised machine learning algorithm based on ensemble learning. Ensemble learning is a type of learning where you join different types of algorithms or same algorithm multiple times to form a more powerful prediction model. The random forest algorithm combines multiple algorithm of the same type i.e. multiple decision trees, resulting in a forest of trees, hence the name “Random Forest”. The random forest algorithm can be used for both regression and classification tasks. The following are the basic steps involved in performing the random forest algorithm:

- Pick N random records from the dataset.
- Build a decision tree based on these N records.
- Choose the number of trees you want in your algorithm and repeat steps 1 and 2.
- In case of a regression problem, for a new record, each tree in the forest predicts a value for Y (output). The final value can be calculated by taking the average of all the values predicted by all the trees in forest. Or, in case of a classification problem, each tree in the forest predicts the category to which the new record belongs. Finally, the new record is assigned to the category that wins the majority vote.

**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 and this model can yield a high predictability rate. Support Vector Machines are perhaps one of the most popular and talked about machine learning algorithms. They were extremely popular around the time they were developed in the 1990s and continue to be the go-to method for a high-performing algorithm with little tuning.

- How to disentangle the many names used to refer to support vector machines.
- The representation used by SVM when the model is actually stored on disk.
- How a learned SVM model representation can be used to make predictions for new data.
- How to learn an SVM model from training data.
- How to best prepare your data for the SVM algorithm.
- Where you might look to get more information on SVM.
II. Conclusion

The analytical process started from data cleaning and processing, missing value, exploratory analysis and finally model building and evaluation. The best accuracy on public test set is higher accuracy score of Speech for KNN and Tremor for Random Forest algorithm. This brings some of the following insights about diagnose the Parkinson disease. Early diagnosis of Parkinson’s is most important for the patient to reduce its impact. To presented a prediction model with the aid of artificial intelligence to improve over human accuracy and provide with the scope of early detection. With our proposed prediction model we aim to make it easier for doctors to do precise diagnosis and prediction of PD, both of which have human limitations due to the method of detection of PD that is used now. It can be inferred from this model that, area analysis and use of machine learning technique is useful in developing prediction models that can help a doctor reduce the long process of diagnosis and eradicate any human error. To separate the work of detection and prediction methods to detect and measure the area of brain that is affected due to PD and use that data in machine learning to create the prediction model with accuracy is higher comparing other models.

FUTURE WORK

- Hospitals want to automate the detecting the disease persons from eligibility process (real time) based on the account detail.
- To automate this process by show the prediction result in web application or desktop application.
- To optimize the work to implement in Artificial Intelligence environment.

REFERENCES