Unsupervised Learning for clustering the disease stages in Parkinson's

BADRINARAYANAN SRINIVASAN

2011

DEGREE PROJECT
Computer Engineering

Programme		Reg number	Extent
Master's Programme in Computer Engineering - Applied Artificial Intelligence		E4007D	15 ECTS
Name of student		Year-Month-Day	
BADRINARAYANAN SRINIVASAN		2011-04-12	
Supervisor		Examiner	
Mark Dougherty		Hasan Fleyeh	
Company/Department		Supervisor at the Company/Department	
Department of Computer Engineering, Dalarna University		Mark Dougherty	
Title			
Unsupervised learning for clustering the disease stages in Parkinson			
Keywords			
Hoehn and yaar, Unsupervised Learning, Parkinson Disease, Artificial Intelligence, Neural Network, Data mining, Computer Engineering, PCA, wavelets			

Abstract

Parkinson's disease (PD) is the second most common neurodegenerative disorder (after Alzheimer's disease) and directly affects upto 5 million people worldwide. The stages (Hoehn and Yaar) of disease has been predicted by many methods which will be helpful for the doctors to give the dosage according to it. So these methods were brought up based on the data set which includes about seventy patients at nine clinics in Sweden. The purpose of the work is to analyze unsupervised technique with supervised neural network techniques in order to make sure the collected data sets are reliable to make decisions. The data which is available was preprocessed before calculating the features of it. One of the complex and efficient feature called wavelets has been calculated to present the data set to the network. The dimension of the final feature set has been reduced using principle component analysis. For unsupervised learning k-means gives the closer result around 76% while comparing with supervised techniques. Back propagation and J4 has been used as supervised model to classify the stages of Parkinson's disease where back propagation gives the variance percentage of 76-82%. The results of both these models have been analyzed. This proves that the data which are collected are reliable to predict the disease stages in Parkinson's disease.

ACKNOWLEDGMENT

I owe a great many thanks to a great many people who helped me and supported me during this work.

Professor Mark Dougherty, the supervisor of the thesis who helped me to understand the concepts and I always inspired by him for his unique thoughts.

I also would like to thank my other teachers who helped me a lot. Especially PHD student Melvudin Memedi who gave me lots of thoughts and ideas to make the work perfect.

Last but not least, I am always thankful for my parents and friends who always have trust on me and supported me in different situations.

Table of contents

7. Conclusion -- **29**

References

LIST OF FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION:

Parkinson's disease (PD) is the second most common neurodegenerative disorder (after Alzheimer's disease) and directly affects up to 5 million people worldwide [1]. PD is a disorder of the brain that leads to shaking (tremors) and difficulty with walking, movement, and coordination [2]. PD is both a chronic and progressive disease. Chronic means that it lasts for long time. Progressive means that symptoms get worse over time. The symptoms of the disease will vary among individuals and also the speed at which the disease progresses differ for each patient.

Hoehn and yahr (1967) described the progression of PD in five stages with increasing disability. In stage 1 the disease was confined to one side of the body (unilateral). In stage 2, the disease was bilateral but there was no problem with balance. In stage 3, postural instability becomes an issue. In stage 4 there is severe disability but the patient is still able to walk or stand. In stage 5, the patient is wheelchair bound or bedridden unless aided [5].

Some of the symptoms for this disease are uncontrollable shaking, stiffness of the muscles, a loss of balance and coordination, depression and so on. More information about the symptoms of this disease can be found Parkinson's disease foundation. There are many medications available to treat the symptoms of Parkinson's, although none yet that actually reverse the effects of the disease.

It is common for people with PD to take a variety of these medications – all at different doses and at different times of day - in order to manage the symptoms of the disease. The classes of medications available to treat Parkinson's on the market include Carbidopa/Levodopa therapy, Dopamine Agonists, Anticholinergics, MAO-B Inhibitors, COMT Inhibitors, Other medications [6].

Can medicines or medications cure PD? The answer for this question is no. There is no cure for this disease however medications will provide some obstacles for the disease progression also some of the symptoms will be relieved. Medications should be carefully done and it must be individually tuned. If the dosing exceeds then the patient will be affected by Dyskinesia. While keeping track of medications can be a challenging task, understanding your medications and sticking to a schedule will provide the greatest benefit from the drugs and avoid unpleasant "off" periods due to missed doses [6].

Developing a machine learning methods to classify the disease stages based on the observation (data set) of a patient would be easier for the doctors to give the appropriate dosage to the patients. There are many efficient models has been developed using neural networks or machine learning methods based on the data set.

1.1 PROBLEM DESCRIPTION:

There are many works which has been defined as a perfect model to find out or classify the disease stages of patients in PD using Machine learning methods or neural networks. Machine learning methods are completely reliable on the data set. Based on the data set the model will be trained and which will be used to predict the disease stages. What if the data which are collected are not reliable but still the data set can be trained using a model. If the prediction leads to a wrong classification then the patient will get a wrong dosage which is more complicated.

1.2 OBJECTIVE:

The main objective of the work is to find out the reliability of the data patterns which have been collected. If the data which are collected are not reliable but still the data set can be trained using a model. If the data patterns are wrongly classified as disease stages then the patient will get a wrong dosage and this will lead to new symptoms.

In order to avoid this problem one has to check the reliability of the data patterns. Reliability can be checked using neural network techniques. The solution has been briefly explained in proposed solution.

1.3 RELATED WORK:

There are several AI techniques that have been tried for several purposes in Parkinson disease. There are few listed below.

Some of the related work about using AI techniques to aid in PD diagnosis and other types of mental disorder are neural network analysis of learning in autism. Neural Networks and Psychopathology by Cohen [9], A model of attention impairments in autism by Bjorn and Balkenius [10], An artificial neural network stimulating performance of normal subjects and schizophrenics on the Wisconsin Card Sorting Test by Berdia and Metz [11], Recognition of ongoing mental activity with artificial neural network by Ivanitsky and Naumov [12] and Online prediction of self-paced hand-movements from sub thalamic activity using neural networks in Parkinson's disease by Loukas and Brown [13].

1.4 PROPOSED SOLUTION:

As we see the problem is complicated because of the data reliability, in this work the reliability has been analyzed with the help of neural network techniques. The solution which has been proposed is cluster the data patterns using unsupervised learning techniques and compares the result after classifying the disease stages using supervised learning techniques. If the results of both these models are closer to each other then it shows that the data are reliable in order to predict the stages if there is a gap between these results then the data collection has to be sorted out with the help of experts.

In order to apply these techniques, features have been calculated for the pre-processed data. The features which will present the entire data set are mean, standard deviation and the most important new feature called wavelets.

Discrete wavelet transform will be calculated for the entire data set. Once all the features are ready then by using the necessary features will be applied into the network. The important features will be extracted using principle component analysis.

Chapter 2

THEORETICAL BACKGROUND:

Forty years after its discovery, levodopa remains the most effective medication for Parkinson's disease. In fact, 70 to 80 percent of treated Parkinson's patients are on levodopa therapy. Levodopa is the "gold standard" by which all treatments for Parkinson's are measured [6].
Observing the patient's progress is very important process of medication. Single or a few observations are not reliable in order to assess how much of the time the patient spends in the different states (off, on or dyskinetic), and how much state varies.

2.1 Personal Digital assistant:

A test battery was constructed and implemented on a hand computer with a touch screen and built-in mobile communication to be used by patients at home as a telemedicine approach to observe their progress. The test battery consists of a combined e-diary with Parkinson-related questions and on-screen motor tests (different tapping tests and spiral drawings). Around 65 patients data are available with between one and six weekly tests periods each available. In fluctuating patients the test battery should typically be used several times daily in the home environment, over periods of about one week. The aim of this test battery is to provide status information in order to evaluate treatment effects in clinical practice and research, follow up treatments and disease progression and predict outcome to optimize treatment strategy [3].
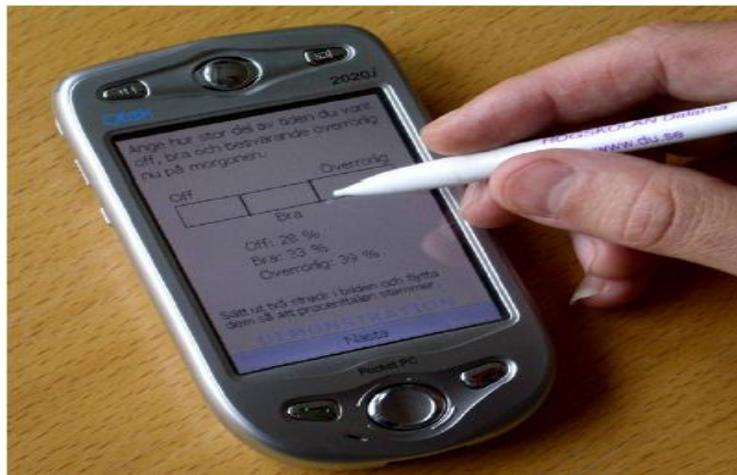
Figure 2.1 – A photo of a Qtek 2020i Pocket PC, running the test battery application [3].

Patients enter e-diary data and perform motor test in the PDA, and the data is transmitted to a database server via a mobile net. The database server receives the data from the PDA, for further processing and storage, to be displayed on demand from various users. An existing system parses the data from XML files, inserts the data into a database as a raw data and distributes it into different tables [7].

The use of classifier systems in medical diagnosis is increasing gradually. Recent advances in the field of artificial intelligence have led to the emergence of expert systems and Decision Support Systems (DSS) for medical applications. Moreover, in the last few decades computational tools have been designed to improve the experiences and abilities of doctors and medical specialists in making decisions about their patients. Without doubt the evaluation of data taken from patients and decisions of experts are still the most important factors in diagnosis. However, expert systems and different Artificial Intelligence (AI) techniques for classification have the potential of being good supportive tools for the expert. Classification systems can help in increasing accuracy and reliability of diagnoses and minimizing possible errors, as well as making the diagnoses more time efficient [8].

Let us see the detailed description about hoehn and yaar stages in Parkinson's disease.

Hoehn and Yahr Staging of Parkinson's disease

1. Stage One
    a. Signs and symptoms on one side only
    b. Symptoms mild
    c. Symptoms inconvenient but not disabling
    d. Usually presents with tremor of one limb
    e. Friends have noticed changes in posture, locomotion and facial expression
2. Stage Two
    a. Symptoms are bilateral
    b. Minimal disability
    c. Posture and gait affected
3. Stage Three
    a. Significant slowing of body movements
    b. Early impairment of equilibrium on walking or standing
    c. Generalized dysfunction that is moderately severe
4. Stage Four
    a. Severe symptoms
    b. Can still walk to a limited extent
    c. Rigidity and bradykinesia
    d. No longer able to live alone
    e. Tremor may be less than earlier stages
5. Stage Five
    a. Cache tic stage
    b. Invalidism complete
    c. Cannot stand or walk
    d. Requires constant nursing care

This rating system has been largely supplanted by the Unified Parkinson's Disease Rating Scale, which is much more complicated [5].

2.2 SIGNAL PROCESSING:

Signal processing treats signals as stochastic processes, dealing with their statistical properties (e.g., mean, covariance, etc.). In many areas signals are modelled as functions consisting of both deterministic and stochastic components.

2.2.1 Why Discrete wavelet Transform?

In previous work Fast Fourier transform has been used to classify the disease stages in PD. However in general Fourier transform can show the frequency but not the time. In other words the data set is non stationary and the wavelets are efficient to work with non stationary data. There are time slot ids with every 4 hours of interval. However the patient doesn't take the test 24 (p.m) and 4 (a.m) since it's obvious he will be sleeping. Why do we need the data at these intervals? We have the data at the intervals 8, 12, 16 and 20 but if we get the answers for all questions at remaining intervals it would make the prediction more efficient. In previous work interpolation has been done to replace the intervals 24 and 4 but it may leads to wrong prediction of the disease stages. So in this work in addition to mean and standard deviation wavelets has been calculated which will be presented as time series graph. Let us see some more information about wavelet transform.

The idea behind these time-frequency joint representations is to cut the signal of interest into several parts and then analyze the parts separately. It is clear that analyzing a signal this way will give more information about the when and where of different frequency components, but it leads to a fundamental problem as well: how to cut the signal? Suppose that we want to know exactly all the frequency components present at a certain moment in time.

The wavelet transform or wavelet analysis is probably the most recent solution to overcome the shortcomings of the Fourier transform. In wavelet analysis the use of a fully scalable modulated window solves the signal-cutting problem. The window is shifted along the signal and for every position the spectrum is calculated. Then this process is repeated many times with a slightly shorter (or longer) window for every new cycle.

In the end the result will be a collection of time-frequency representations of the signal, all with different resolutions. Because of this collection of representations we can speak of a multiresolution analysis. In the case of wavelets we normally do not speak about time-frequency representations but about time-scale representations, scale being in a way the opposite of frequency, because the term frequency is reserved for the Fourier transform. This is the main reason to choose discrete wavelet transform when compared to other dsp methods.

There are two most used discrete wavelet transform namely
1. Haar and
2. Daubechies

Haar is the basic or default transforms which can also be called as Daubachies 1. Let us see some more details about haar and Daubachies.

2.2.2 Haar wavelet Transform:

The Haar wavelet is a certain sequence of rescaled square-shaped functions which together form a wavelet family or basis. Wavelet analysis is similar to Fourier analysis in that it allows a target function over an interval to be represented in terms of an ortho-normal function basis. The Haar sequence is now recognised as the first known wavelet basis and extensively used as a teaching example in the theory of wavelets.

Using the discrete wavelet transform, one can transform any sequence $(a_0, a_1, \ldots, a_{2n}, a_{2n+1})$ of even length into a sequence of two-component-vectors $((a_0, a_1), \ldots, (a_{2n}, a_{2n+1}))$.

If one right-multiplies each vector with the matrix H_2, one gets the result $((s_0, d_0), \ldots, (s_n, d_n))$ of one stage of the fast Haar-wavelet transform.
Usually one separates the sequences s and d continues with transforming the sequence s. The Haar transform is the simplest of the wavelet transforms. This transform cross multiplies a function against the Haar wavelet with various shifts and stretches, like the Fourier transform cross-multiplies a function against a sine wave with two phases and many stretches.

2.3.3 Daubechies wavelet Transform:

The Daubechies wavelets are a family of orthogonal wavelets defining a discrete wavelet transform and characterized by a maximal number of vanishing moments for some given support. With each wavelet type of this class, there is a scaling function (also called father wavelet) which generates an orthogonal multi resolution analysis. In general the Daubechies wavelets are chosen to have the highest number A of vanishing moments, (this does not imply the best smoothness) for given support width N=2A, and among the 2^{A-1} possible solutions the one is chosen whose scaling filter has external phase.

The wavelet transform is also easy to put into practice using the fast wavelet transform. Daubechies wavelets are widely used in solving a broad range of problems, e.g. self-similarity properties of a signal or fractal problems, signal discontinuities, etc. The Daubechies wavelets are not defined in terms of the resulting scaling and wavelet functions; in fact, they are not possible to write down in closed form.

2.3 PRINCIPLE COMPONENT ANALYSIS:

We will see detailed background of principle component analysis and in later part we will see how and where it has been used in this work.

Principle component analysis is the oldest and most used technique to reduce the dimensions or extracting most needed features. Principle component analysis is a method to extract the important features from several features. The features which have been extracted are the linear combinations of several features. It is useful when we have obtained data on a number of variables and believe that there is some redundancy in those variables.

In this case, redundancy means that some of the variables are correlated with one another, possibly because they are measuring the same construct. Because of this redundancy, we believe that it should be possible to reduce the observed variables into a smaller number of principal components. The number of principal components extracted is equal to the number of observed values being analyzed. Only the first few principal components account for meaningful amounts of variance, so only the first principal components are interpreted and used.

The first component extracted in a principal component analysis accounts for a maximal amount of total variance in the observed variables; this means that the first component will be correlated with some of the observed variables. The second component extracted will account the amount of total variance in the data set that was not accounted by the first component, and it will be uncorrelated with the first component.

The remaining components account for a maximal amount of variance in the observed variables that was not accounted by the preceding components, and are uncorrelated with preceding components, this is why the only the first few components are usually retained and interpreted. When using PCA the observed variables are standardized in the course of the analysis, it means that each variable is transformed so that it has a mean of zero and a variance of one. The total variance in the data set is simply the sum of the variance of these observed variables [16].

2.4 BACK PROPAGATION:

Back propagation is the generalization of the Widrow-Hoff learning rule to multiple-layer networks and nonlinear differentiable transfer functions. Input vectors and the corresponding target vectors are used to train a network until it can approximate a function, associate input vectors with specific output vectors, or classify input vectors in an appropriate way. Networks with biases, a sigmoid layer, and a linear output layer are capable of approximating any function with a finite number of discontinuities.

Standard back propagation is a gradient descent algorithm, as is the Widrow-Hoff learning rule, in which the network weights are moved along the negative of the gradient of the performance function. The term back propagation refers to the manner in which the gradient is computed for nonlinear multilayer networks.

Properly trained back propagation networks tend to give reasonable answers when presented with inputs that they have never seen. Typically, a new input leads to an output similar to the correct output for input vectors used in training that are similar to the new input being presented. This generalization property makes it possible to train a network on a representative set of input/target pairs and get good results without training the network on all possible input/output pairs [19].

2.4.1 Architecture of multi layer Perceptron:

It's good to know about the multi layer Perceptron in order to understand the algorithm. The below diagram shows multi layer Perceptron with
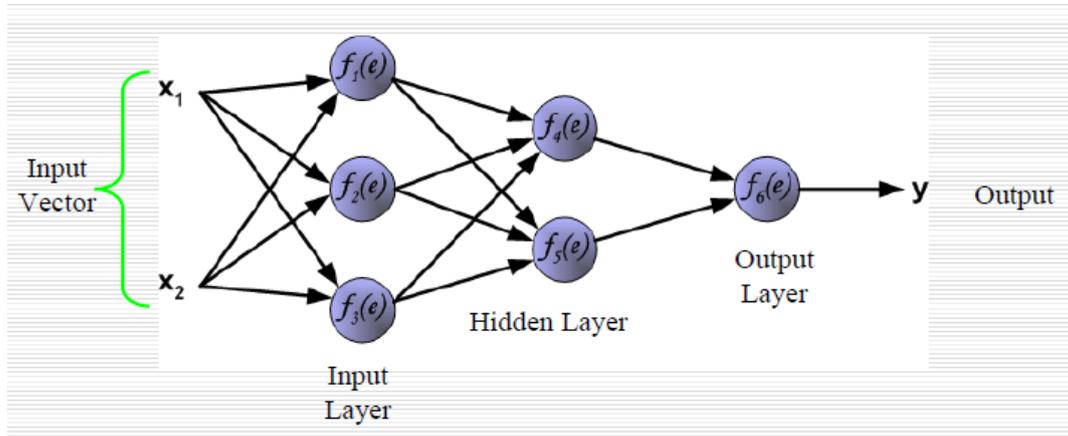

Figure 2.2 Multi Layer Perceptron [20]

- Input layer
- Hidden layer
- Output layer
- Input and output vector

2.4.2 Learning:
Error back-propagation learning consists of two passes through the different layers of the network:

2.4.2.1 Forward Pass:
1. An input vector is applied to the sensory nodes of the network.
2. Its effect propagates through the network layer by layer.
3. A set of outputs is produced as the actual response of the network.
4. During this pass the synaptic weights of the networks are all fixed.

2.4.2.2 Backward Pass:
1. An error signal is calculated and propagated backward through the network against the direction of synaptic connections.
2. The synaptic weights are adjusted to make the actual response of the network move closer to the desired response in a statistical sense.
3. During this pass, the synaptic weights are all adjusted in accordance with an error-correction rule.

2.4.3 Types of input signals:

Function signal
1. It is the input signal that comes in at the input end of the network

2. It propagates forward through the network, and emerges at the output end of the network as an output signal.

Error signal

1. An error signal originates at an output neuron of the network
2. It propagates backward through the network.
3. It is called as "error signal" because its computation by every neuron of the network involves an error-dependent function.

2.5 K-MEANS:

K-means is an efficient and simplest algorithm which can be used for clustering. The basic step of k-means clustering is simple. In the beginning we determine number of cluster K and we assume the centroid or center of these clusters. We can take any random objects as the initial centroids or the first K objects in sequence can also serve as the initial centroids.

Then the K means algorithm will do the three steps below until convergence

Iterate until stable (= no object move group):

1. Determine the centroid coordinate
2. Determine the distance of each object to the centroids
3. Group the object based on minimum distance

Chapter 3

METHODOLOGY

Let us see the methodology of the work. The study progresses over three years and includes about 70 patients at about nine clinics in Sweden. During the first year, the test battery is used one week every third month, and during year two and three, one week every six month. The patients respond to questions and perform tests in the home environment four times daily during test weeks [18].

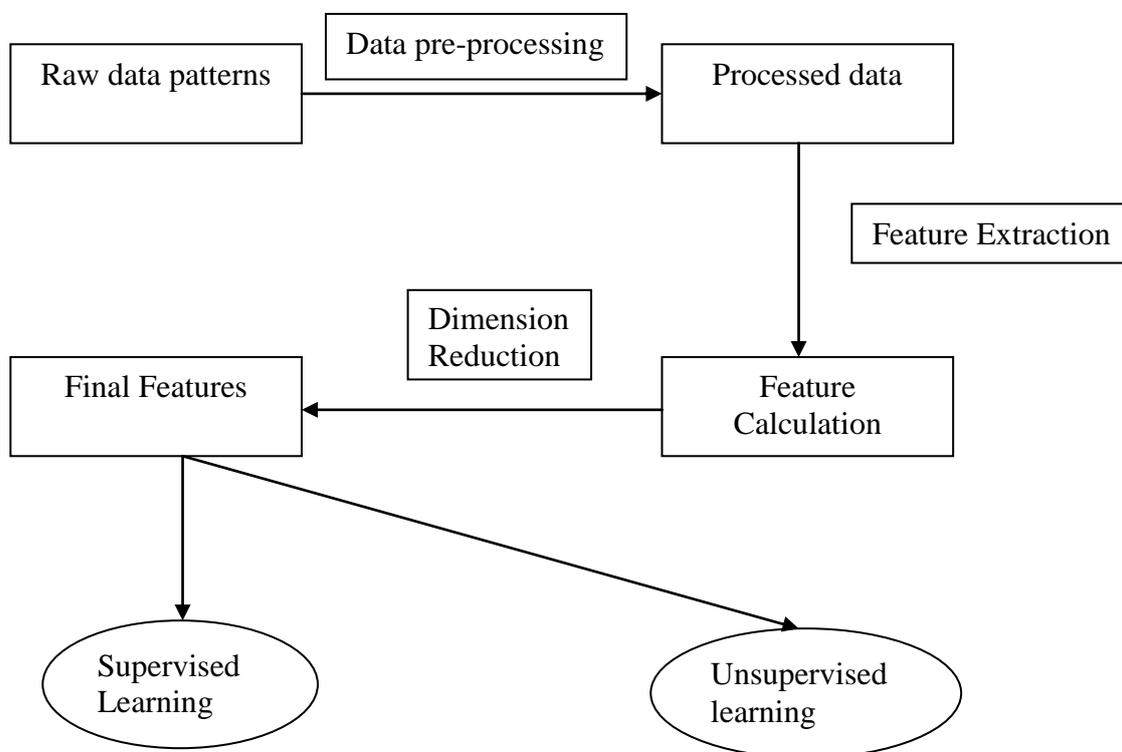The methodology has been shown in the block diagram in order to see how the work process takes place.

Figure 3.1 Block Diagram

3.1 DATA ACQUISTION:

Evaluating status in patients with motor fluctuations is complex and occasional observations/measurements do not give an adequate picture of time spent in different states [17]. In measuring the health effects, a hand computer based test battery with diary questions and motor test will be used.
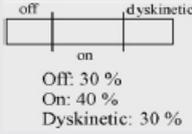
It is important to know the questions that have been answered by the patients. Understanding the data is important process and it can be said as data analysis. A database is available with around 300 weekly registration periods with about four registrations per day.

Summary scores from the spirals based on wavelet transform methods have already been computed. Patients own assessments of symptoms, speed and accuracy in tapping tests are available in the database. Around 200 of these registration periods, Hoehn and Yahr disease stages are assessed by medical doctors.

3.1.1 Data Set:

The database which was discussed previously was collected from several patients. The database contains 9 tables out of which 6 tables are related to finding out the stages of Hoehn and yahr. The table (3.1 & 3.2) shown below is overall summary of the questions.

Table 3.1 – Question Definitions [3]

Question #	Text	Answer alternatives
1	Have you had difficulty to walk (ca 100 m) during the last 4 h/this morning?	Unable to walk Difficult With effort Fairly well Walks well
2	Mark the portion of time you have been off, on and dyskinetic, during the last 4 h/this morning Tap to place two lines in the figure and move them to adjust percentages[a]	off dyskinetic on Off: 30 % On: 40 % Dyskinetic: 30 %
3	How off have you been at worst during the last 4 h/this morning?	Extremely off Quite a bit Moderately Slightly Not at all off
4	How dyskinetic have you been at worst during the last 4 h/this morning?	Extremely dyskinetic Quite a bit Moderately Slightly Not at all dyskinetic
5	Have you had painful cramps (dystonia) during the last 4 h/this morning?	All the time Most of the time Half the time A little of the time No cramps at all
6	Have you been satisfied with your function during the last 4 h/this morning?	Not at all satisfied Slightly Moderately Quite a bit Completely satisfied
7	How is your condition right now?	Very dyskinetic Moderately dyskinetic A little dyskinetic On A little off Moderately off Very off

In this question, the patients were instructed to divide the time spent in "off", "on" and "dyskinetic" states by placing two vertical bars in a rectangle, giving three areas with sizes proportional to the time spent in these states.

Table 3.2 – Definitions of motor test [3]

Test #	Test description	Test area
8	Alternately tapping two fields (un-cued) using right hand. Two active buttons (red) are displayed. The buttons both become inactive (black) 20 s after the first button was pressed.	
9	Same as 8 but using left hand.	
10	Two buttons, where one is active (red) and the second is inactive are displayed. The buttons alternately become active with increasing speed. Alternations start with 2 s per tap and speed increases with constant acceleration to 0.5 s per tap after 20 s. A sound is heard and a small filled circle is briefly (timeout) shown after a tap. Sounds on correct and missed taps are different. Dominant hand is used.	
11	Tapping an active field that changes location when tapped, Four buttons, where one is active (red) and the other three are inactive, are displayed. When an active button is tapped, one (randomly selected) of the other three buttons becomes active. Cueing is supplied as in test 10 above. Dominant hand is used.	
12–14	Tracing a pre-drawn Archimedes spiral from the centre and out using dominant hand without supporting it. The test is repeated three times and patients are instructed that it should be completed in approximately 10 s per drawing.	

3.1.2 Diary Questions:

The diary questions and their answer alternatives are given in Table 3.1. Questions 1–6 relate to the previous 4h, or 'this morning', depending on the actual time-of-day. Question 7 relates to right now and allows seven steps from −3 (very off) to +3 (very dyskinetic). Question 2 gives three answers in percent, whereas questions 1, 3, 4, 5 and 6 are of verbal descriptive scale type between 1 (worst) and 5 (best) [3].

3.1.3 Motor Test:

Detailed descriptions and illustrations of the user interface of each motor test are given in Table 3.2. The tapping tests (Table 3.2, #8–11) display square tapping and the questions 12 – 14 for spiral test.

3.1.4 about the dataset:

In the above section the details about the questionnaire has been explained. Now let us see how the information was stored in a data base. This section will show the clear understanding about the data set. As you see from table 3.1 and 3.2, the total numbers of questions are 14 and these questions are separated into few sections while storing it in the database as tables. The day and time at which patients answered has been noted though there is a column called time slot id which are 8, 12, 16 and 20. There is one more column called period Id which says the month the patient took the test. The periods are from month 0 to month 18. Month 0 called as baseline. Since the database is in the form of tables SQL has been used in order to remove the noises and made some modifications while using it.

3.2 FEATURE EXTRACTION:

Data preprocessing is another important step because real world data are generally Incomplete (lacking attribute values, lacking certain attributes of interest), or containing only aggregate data, Noisy (containing errors or outliers), Inconsistent containing discrepancies in codes or names. In this case the data are collected using PDA so most of the errors have been avoided. However the data stored in the database has some errors in it.

1. There are some missing values in the dataset which has been replaced by the mean of the respective attribute or variable.
2. Some of the patients didn't answer any questions for particular period so they are completely removed from the final input data set**.**
3. **Pooling:** The actual output of few patterns has been modified in order to maintain the result closer. The patient id E0202_month6 has been changed from 1.5 to 2 and the patient who have stages 0 and 1 are removed. Because the patient who have the stage 0 and 1 are very few inexact numbers 1. So in the training process the classifier may not converge due to less number of patterns which has those stages.

Before calculating the features of the dataset the missing values or errors have been replaced with the mean. The total number of records or rows in the final dataset is 9593. Now let us calculate the features for this dataset based on each questions, period and the patient id.

3.2.1 Features:

What are the features calculated for this work? We are going to see in detail about this. But why these features playing significant role? In simple words, a feature summarizes or describes the collection of data set using statistical methods. The most common features which you can see often are mean, median, standard deviation and so on.

Which feature should be calculated? It's a tricky question. It's completely problem dependent. For this work the most important features that have been calculated are Wavelets, Standard deviation and mean. Let us see in detail.

3.2.1.1 Mean:

For a data set, the mean is the sum of the values divided by the number of values. The general formula for mean is shown below:

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

Where $\bar{x}$ the mean, n is the number of values or samples of x

X: axis – period id

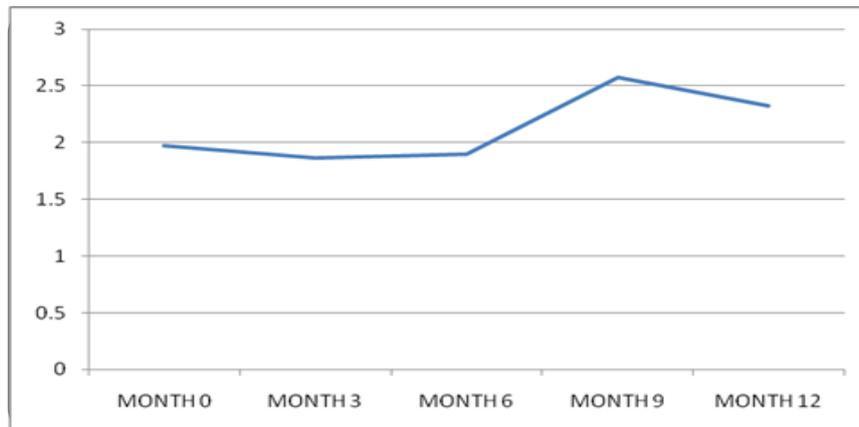Y: axis – mean values for question 1for the patient id E0301

Figure 3.2 Mean for question 1

Fig. 3 shows the mean for question 1. As you can see the most of the patients answered for question 1 between 3 and 4. Very few people answered 1. Here the value of n is 220 i.e., the total number of patterns are 220. So we can conclude many things from this graph. So we have calculated mean for all the questions and stored in a table for further use.

3.2.1.2 Standard Deviation:

Standard deviation is a widely used measurement of variability or diversity. It shows how much variation or dispersion there is from the mean. A low standard deviation indicates that the data points tend to be very close to the mean, whereas high standard deviation indicates that the data are spread out over a large range of values. The general formula for standard deviation is

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2},$$

Where s is standard deviation, n is the total number of values of x and $\overline{x}$ is the mean.

X:axis – Period id
Y:axis – standard deviation values for question 1 of the patient E0301
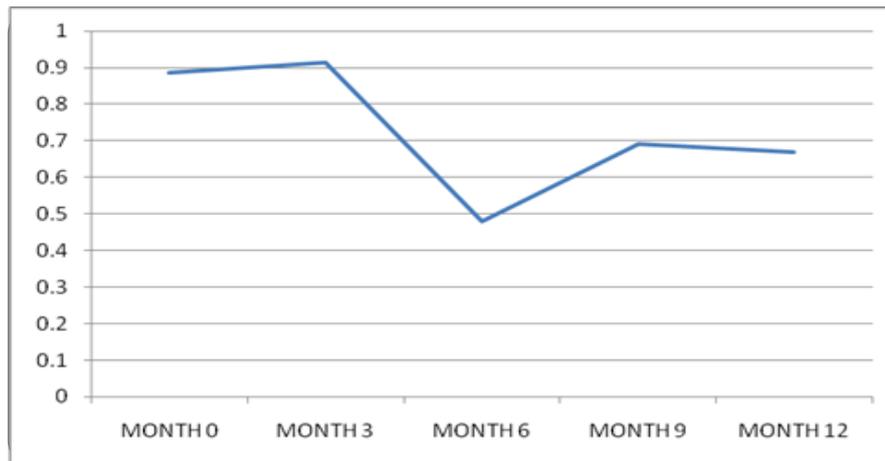
Figure 3.3 Standard deviation for question 1

Fig. 4.2 shows the standard deviation for the question 1. As you can see the values are higher so the deviation is more from the mean.

3.2.1.3 Wavelets:

Calculating the wavelets is complicated process for the larger data set. There are plenty of calculations has been done thanks to the MATLAB which was very useful to do it in shorter time. In order to calculate the discrete wavelet transform one has to choose which kind of DWT to use either HAAR or DAUBACHIES. However choosing a transformation is depends on each application. Here both these transformation has been calculated (haar and db2).

The general formula to calculate the discrete wavelet transform is

[cA,cD] = dwt(X,'wname','mode',MODE)

Where cA is coefficient of approximation

Wname is the wavelet name (HAAR and Daubechies), X is the input vector.

3.2.1.4 Border Distortion:

Before calculating the transform, the input data set is maintained as 28 records for each question and for each patient and period. If the record has more than 28 records then the first 28 were chosen. If the record doesn't have 28 records then, padding the original data set to make it 28.The dwt mode in MATLAB sets the signal or image extension mode for discrete wavelet and wavelet packet transforms. The extension modes represent different ways of handling the problem of border distortion in signal and image analysis.

Here to deal with this border distortion problem symmetric padding mode has been used. For instance if the data set has 27 values then the first value of the data set will be placed as 28th element.

Let us see the graphs of original signal and the transformed signal. The original signal shown in the figure 3.3.1

X: axis − Time at which the question answered
Y: axis − Value of question 1
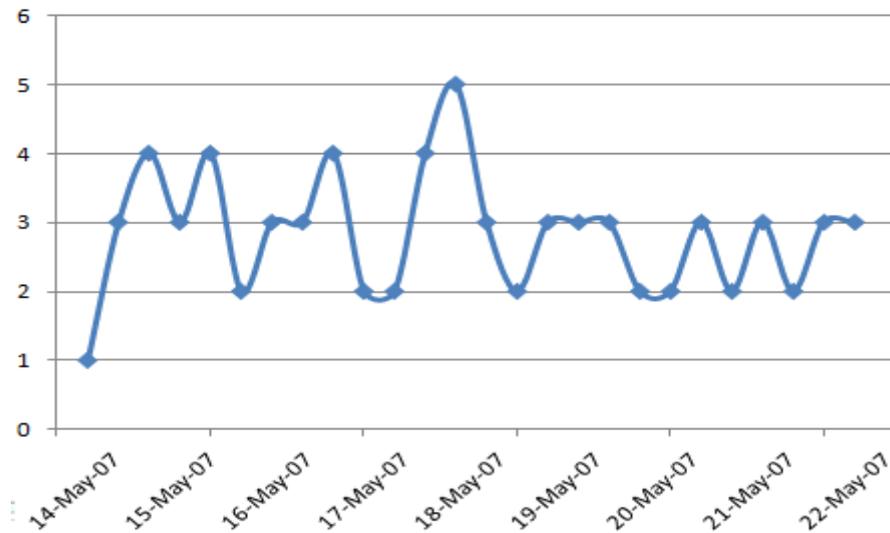The time series data plotted here is belongs to patient id E0506 for question 1 at month 12.
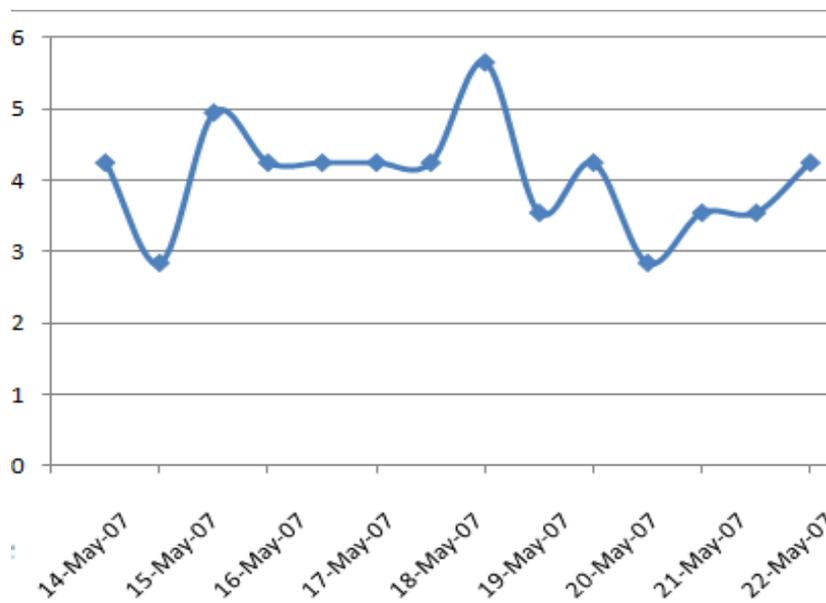

Figure 3.3.1 Original Signal

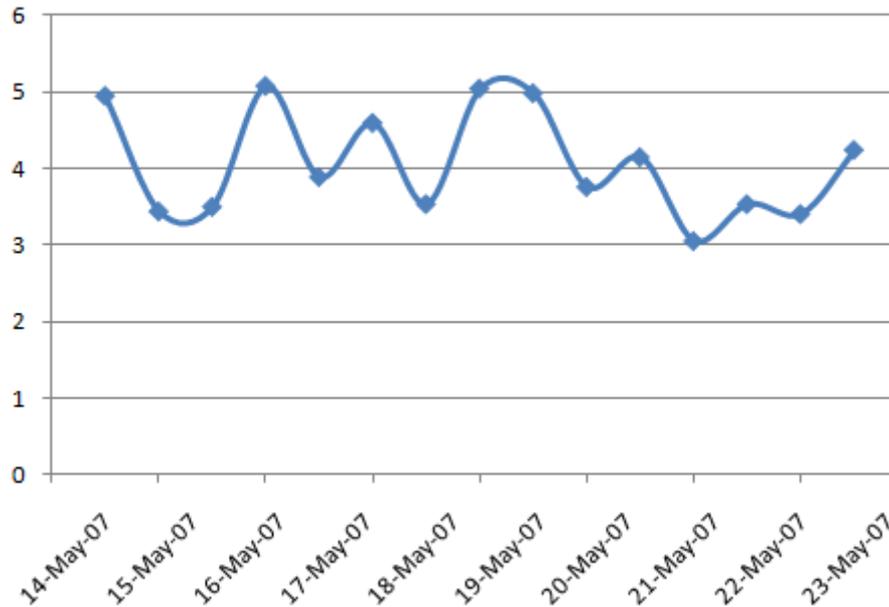**Figure 3.3.2 Approx. coeff. For Haar**

Figure 3.3.3 Approx. coeff. For db2

As you can see the above graphs for haar and db2 plotted shows the wavelet transform of the original signal. The calculation of Coefficient approximation is nothing but the average values of consecutive adjacent values.

These signals can be transformed again and again to make it smoother however it is more complicated to do it. Here the process is just applied one time which shows the half of the total number of data set. Again this 14 data set can be applied to wavelet transform and make it to 7 this will be smoother than the one we see above.

3.3 DIMESNION REDUCTION:

PCA generates a new table with the same number of variables, called the principal components. Each principal component is a linear transformation of the entire original data set. The coefficients of the principal components are calculated so that the first principal component contains the maximum variance (which we may tentatively think of as the "maximum information"). The second principal component is calculated to have the second most variance, and, importantly, is uncorrelated (in a linear sense) with the first principal component. Further principal components, if there are any, exhibit decreasing variance and are uncorrelated with all other principal components.

PCA is completely reversible (the original data may be recovered exactly from the principal components), making it a versatile tool, useful for data reduction, noise rejection, visualization and data compression among other things. The first step in PCA is to standardize the data. Here, "standardization" means subtracting the sample mean from each observation, then dividing by the sample standard deviation. This centres and scales the data.

This calculation can also be carried out using the Z-score function from the Statistics Toolbox. Calculating the coefficients of the principal components and their respective variances is done by finding the Eigen functions of the sample covariance matrix. The matrix V contains the coefficients for the principal components. The diagonal elements of D store the variance of the respective principal components. The coefficients and respective variances of the principal components could also be found using the Princomp function from the Statistics Toolbox.

All these features are gathered and stored in a single input file. The final input files will have mean, standard deviation and wavelets (after performing PCA).

3.4 ATTRIBUTE SELECTION:

Attribute or Feature selection is a fundamental problem in many different areas. Real time problems has many features, however, few features can be avoided in case if it will decay the results. So selecting the features is important step as well. There are several possible ways to apply feature selection. Feature selection algorithms typically fall into two categories: feature ranking and subset selection. Feature ranking ranks the features by a metric and eliminates all features that do not achieve an adequate score. Subset selection searches the set of possible features for the optimal subset. Here there are two kinds of approaches are used.

For unsupervised learning method principle component analysis is used again in order to extract the very important features to implicate the results. As we know principle components gives the linear combination of several attributes.

After applying PCA on the final input data set, the total numbers of features are 16 which are selected based on the threshold value. All these 16 features together represent 85% based on variance. The process of applying PCA explained as follows:

Most often, the first step in PCA is to standardize the data. Here, standardization means subtracting the sample mean from each observation, then dividing by the sample standard deviation. This centers and scales the data. Sometimes there are good reasons for modifying or not performing this step.

This calculation can be carried out using the zscore function from the Statistics toolbox

Y=zscore(x)

Calculating the coefficients of the principal components and their respective variances is done by finding the eigen functions of the sample covariance matrix.

PCA squeezes as much information (as measured by variance) as possible into the first principal components. In some cases the number of principal components needed to store the vast majority of variance is shockingly small: a tremendous feat of data manipulation. This

transformation can be performed quickly on contemporary hardware and is invertible, permitting any number of useful applications.

For supervised learning, weka tool has been used. In the tab select attributes there are several number of approaches can be found. However, the approach selected here based on the attributes which is more equal to princimple components. Chi squared eval has been used used. It falls under featuer ranking category.

Chapter 4

IMPLEMENTATION:

Now let us see the experimental details of the work. Now we have the important features that represent the entire data set. Unsupervised and supervised techniques were applied on these features. For unsupervised technique k-means has been applied and for supervised, back propagation and attribute selected classifier were applied. Let us see in detail.

4.1 UNSUPERVISED LEARNING:

4.1.1 K- means:
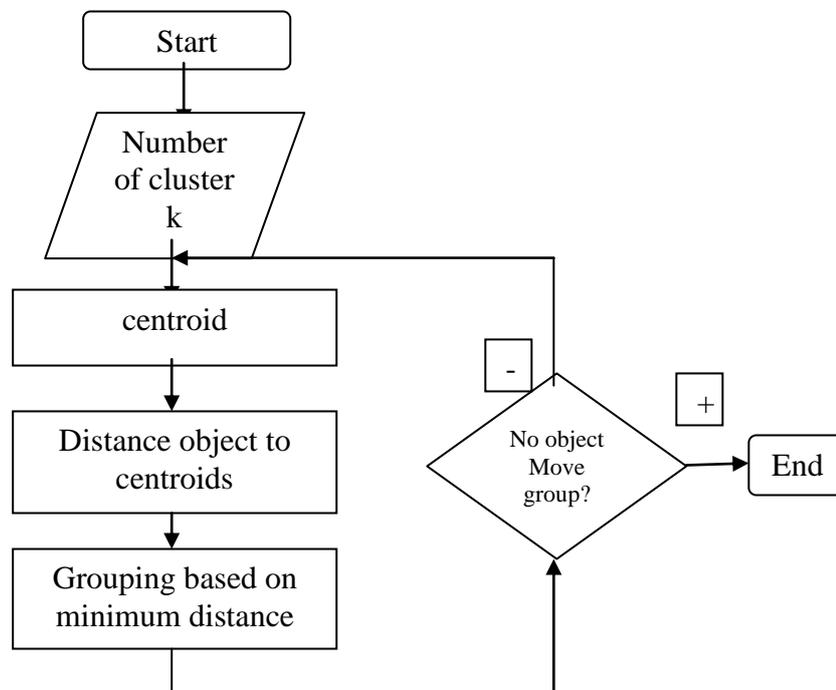
The general algorithm for k means is shown below:

Figure 4.1 K-means Flow chart

K means partitions the points in the N-by-P data matrix X into K clusters. This partition minimizes the sum, over all clusters, of the within-cluster sums of point to cluster centric distances. Rows of X correspond to points, columns correspond to variables. In other words, an algorithm for partitioning (or clustering) N data points into K disjoint subsets S_j containing N_j data points so as to minimize the sum-of-squares criterion.

$$J = \sum_{j=1}^{K} \sum_{n \in S_j} |x_n - \mu_j|^2,$$

Where x_n is a vector representing the nth data point and μ_j is the geometric centroid of the data points in S_j. In general, the algorithm does not achieve a global minimum of J over the assignments. In fact, since the algorithm uses discrete assignment rather than a set of continuous parameters, the "minimum" it reaches cannot even be properly called a local minimum. Despite these limitations, the algorithm is used fairly frequently as a result of its ease of implementation.

The algorithm consists of a simple re-estimation procedure as follows. Initially, the data points are assigned at random to the K sets.

For step 1, the centroid is computed for each set. In step 2, every point is assigned to the cluster whose centroid is closest to that point. These two steps are alternated until a stopping criterion is met, i.e., when there is no further change in the assignment of the data points.

Defining a number of clusters in K-means is tricky part. In order to define that, the analysis about the output stages has to be read clearly. Clear view of the output stages may match the defining the number of clusters. In MATLAB, K-means has been developed as a pre-defined function which makes the job easier.

IDX = KMEANS(X, K)

KMEANS returns an N-by-1 vector IDX containing the cluster indices of each point. By default, KMEANS uses squared Euclidean distances.

X is the input matrix and number of clusters has been defined as 5 (k=5).

4.2 SUPERVISED LEARNING:

4.2.1 J48:

Weka has been used to classify the patterns based on supervised learning. The supervised method used to classify the pattern is attributing selected classifier.

In order to classify the patterns using weka there are few steps that has to be followed:

Step 1: Convert the input file into arff file
Step 2: open the file in weka

Step 3: go to select attributes in weka explorer and press the select attributes tab. Choose the principle components method to evaluate the attributes. The search methods will be prompted as ranker search method.

Step 4: run the select attributes process and you can see the linear combinations of several attributes.

Step 5: Since the ranked attributes are linear combination of several attributes we cannot merge the actual output. So in order to avoid this problem choose the first 15 attributes from the ranked attributes with linear combination and separate it. One can ask will the process give proper result. In order to prove this again the attribute selection process has been ran through Info gain attribute evaluator. The attribute which have been ranked by Info gain and principle components are same.

Step 6: classify the patterns using attribute selected classifier with ranked attributes.

4.2.2 Back Propagation:

Implementing back propagation to classify the stages of Parkinson's disease is a bit tricky task. As discussed earlier the number of features calculated for all the answers is huge. However after applying the feature selection, there are some features has been left out. Only important features which is quite enough for classification has been picked. MATLAB has the tool box functions to implement the back propagation algorithm.

4.2.2.1 Input vector:

The number of input pattern is 220 with 15 features. The vector has been stored in an excel file which can be loaded in MATLAB with the function "xlsread".

Step 1: Load the file

Step 2: Usually the input vector will be standardized however the input vector has been standardized already the next step is to create a network.

Step 3: As an initial guess, ten neurons used in the hidden layer. The resulting network will have five output neurons because the target vectors have five elements (2, 2.5, 3, 4, and 5).

Step 4: The default Levenberg-Marquardt algorithm is used for training

Initially total number of hidden layers assigned as 2 but the results were better when the number of hidden layers increased to 3.

Step 5: Now the network should be trained.

Number of epochs are 15 and the training has stopped after 15 iterations. The results of the back propagation have been explained in the next section briefly.

Chapter 5

RESULT AND ANALYSIS:

As described earlier, the main objective of the work is to find out the reliablity if the data set. In order to find, the processed data patterns has been used in unsupervised and supervised techniques.

5.1 UNSUPERVISED LEARNING K-MEANS:

Number of clusters defined: 5

Total number of patterns: 220

After applying the k means for the patterns, the output clusters has been compared with the actual output and the percentage of classification has been found. 76.33% patterns are correctly clustered into their specific clusters.
 X: axis – silhouette values
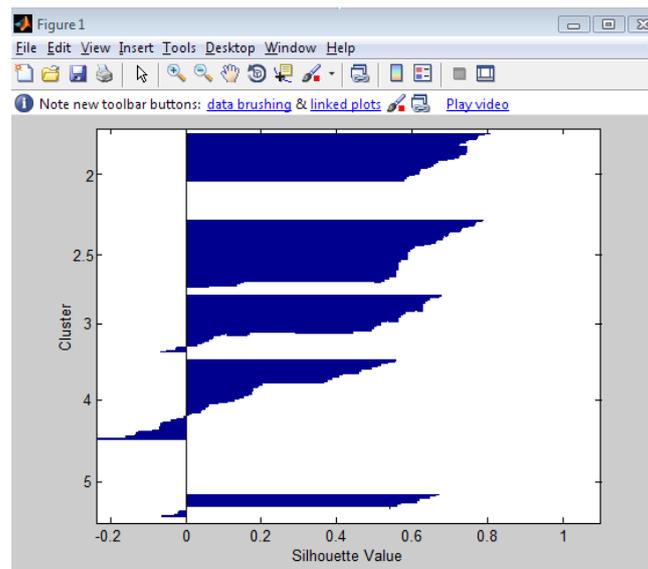Y: axis - clusters

Figure 5.1 Output clusters plotted in scatter graph

Silhouette values - to get an idea of how well-separated the resulting clusters are, you can make a silhouette plot using the cluster indices output from kmeans. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters. This measure ranges from +1, indicating points that are very distant from neighboring clusters, through 0, indicating points that are not distinctly in one cluster or another, to -1, indicating points that are probably assigned to the wrong cluster. silhouette returns these values in its first output.

In the figure 5.1 as you can see all the clusters are almost separated with a larger distance however the points in cluster 4 assigned to wrong one.

From the silhouette plot, you can see that most points in the first, second and third cluster have a large silhouette value, greater than 0.6, indicating that the cluster is somewhat separated from neighboring clusters. However, the fourth cluster contains many points with low silhouette values, and also contains a few points with negative values, indicating that cluster is not well separated.

While checking manually there are few clusters are in wrong clusters. For instance, from the result the pattern 27, 102, 153, 154 should be in the cluster 3 but in the above graph it's in cluster 2. It's because the centroid is closer to both 2 and 3. So these kinds of small variations affect the result in unsupervised method.

5.2 J48 classifier:

As the process explained clearly in the above section let us see the result of it.

After applying the select attributes process the number of attributes has been reduced to 15. The target vector is merged with these ranked attributes and loaded in Weka as you can see in the figure 5.2. In order to train the classifier the patterns has been divided into training set and testing set. This is a tricky part since if you look at the patients who are having the stage 5 is very few when compared to people who are having the stage 4, 3 and 2.5. So one must ensure the test data should contain all the stages of patients in order to train properly and also the test data. The classification result shows that 82% of the test data has been classified properly after training the classifier with train data.

The result of the classification and the confusion matrix are shown below. If you look at the confusion matrix you can notice class 5 has been classified without error and class 4 has 3 errors classified incorrectly and so on. The overall classification for the test data are 82.27%.

```
=== Confusion Matrix ===

 a  b  c  d  e   <-- classified as

 2  1  0  0  0 |  a = 2

 0  9  1  4  0 |  b = 2.5

 0  1 11  0  0 |  c = 3

 0  0  3 11  0 |  d = 4
```

Figure 5.2 J48 confusion matrixes

The patterns have been tested with some other models as well. The results of the model have been showed in the below table. Out of all the other models the above model k means for unsupervised and attribute selected classifier looks impressive.

The decision tree doesn't look promising since the network was over trained by the features. In order to make the work better back propagation has been used since linear classification is not efficient for this data set. Let's see how back propagation shows the result of classification.

5.3 BACK PROPAGATION CLASSIFICATION:

The result of applying back propagation was better than the previous method (attribute selected classifier). Since the previous method was over trained back propagation has been used. The training stops at the right place as you see in the below figure. The training stops at the 15th iteration or epoch. The training has stopped since the error rate reaches .001 which is the default stopping criteria and it has been displayed in the network as well. There are three lines in the plot. The blue one is the training data set and the red one is the test data set. The data set divided into 60% for train the network, 20% to validate the network and 20% for testing the network.

X: axis – Number of epochs or iterations
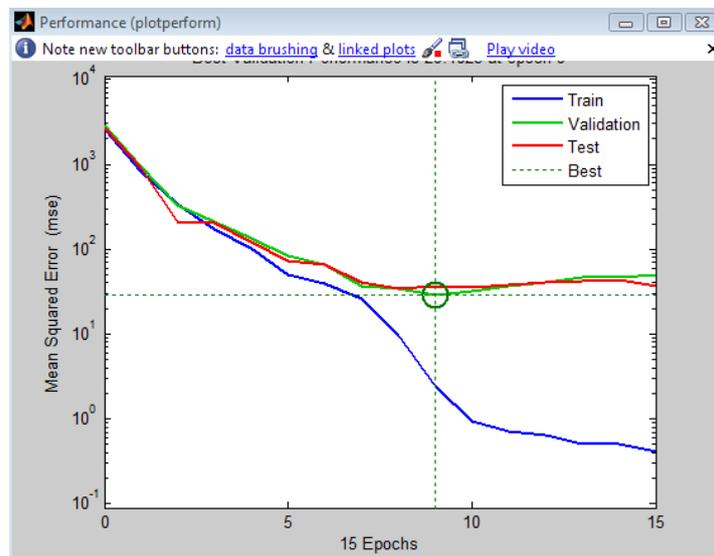Y: axis – Mean squared error values

Figure 5.3 performance plot

The regression plot shows the target data which is classified into respective 5 classes. The graph which is the top left corner is for train data which 60% of the entire data set. The training data set is divided properly since all the classes can be seen in the graph. Because the entire data set has only four patterns which belongs to class 5. The training data has one pattern in it in order to train the network so it can be classified as well. The top right corner graph is plotted for validation data set and bottom left corner graph is plotted for test data and the bottom right corner graph is plotted for the entire data set.

X: axis – actual target classes
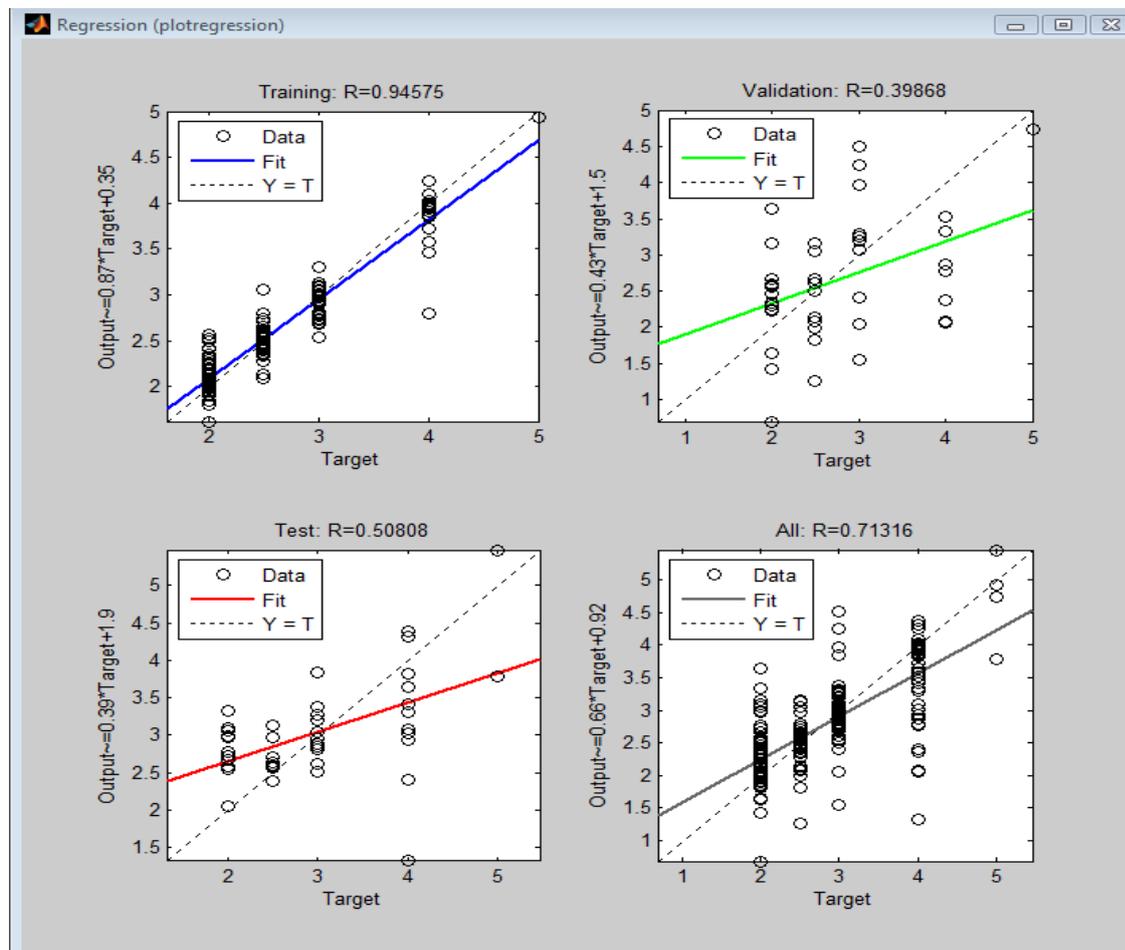Y: axis – desired target classes

Figure 5.4 Regression plot

Since the network assigns different weight at each time it executes we get a different classification percentage. So we executed the network several times and the classification percentage varies between 76-82%. The below figure shows the regression plots of the entire data set which has been executed several times to find out the range of classification.

False positive – 2
False negative – 0

In the test data we added a few false patterns in order to find out the false positives and negatives using the models and after we apply the model we got only 2% of false positives and 0 false negatives.

X: axis – actual target classes
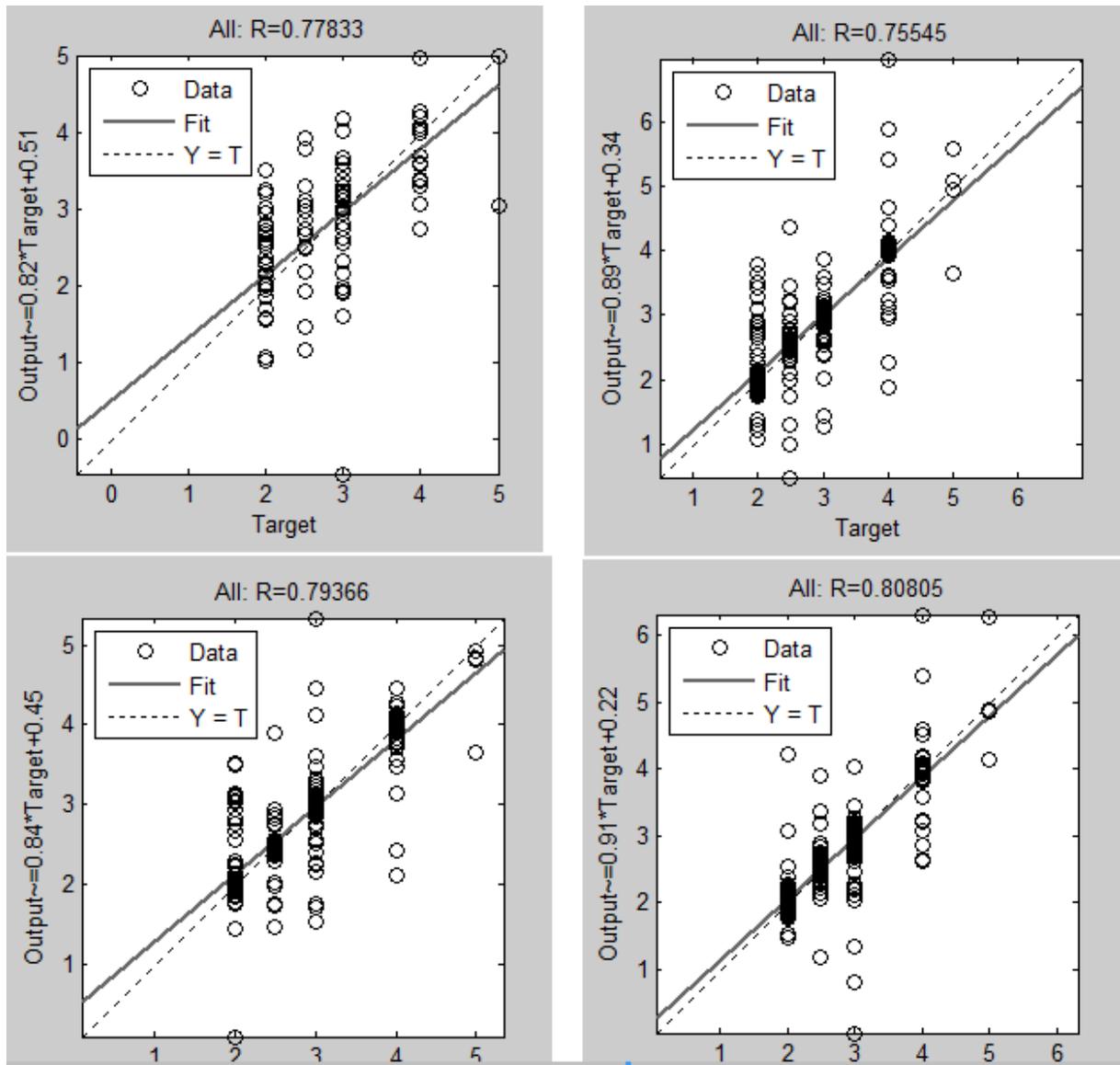Y: axis – desired target classes

Figure 5.5 Regression plot after n executions

So the unsupervised learning model gives 76% accuracy whereas supervised learning model gives classification variance percentage of 76 to 82%. So both these techniques give a closer accuracy to each other so the data patterns which have been collected are reliable to predict the disease stages.

Chapter 6:

CONCLUSION AND FUTURE WORK:

The solution which has been proposed is cluster the data patterns using unsupervised learning techniques and compares the result after classifying the disease stages using supervised learning techniques. If the results of both these models are closer to each other then it shows that the data are reliable in order to predict the stages if there is a gap between these results then the data collection has to be sorted out with the help of experts.

So the models which have been used for unsupervised model cluster the disease stages with the percentage 76% of accuracy. In other words, Hoehn and yaar stages in Parkinson disease has been clustered using unsupervised learning approach. Then using a supervised technique classifies the stages and analyzes the result of these models.

The unsupervised model is able to match the result of supervised learning as closer as possible. As the result discussed in the previous section, k-means gives around 76% of correct classification and Back propagation supervised classification gives around 76-82% of correctly classified test data. Since J48 was over trained back propagation method has been introduced used.

The obtained results showed to be promising by analyzing with supervised classification method. Wavelets are the additional feature added to mean and standard deviation statistical approaches. Haar and db were calculated to cross check both these transformations. Daubachies is the complex transformation when compared to J48. While calculating the principle components of wavelets there were some strong correlations with the other attributes.

The difference between the original signal and the transformed signal clearly shows the important of wavelets. Principle components of the wavelets and the principle components of the overall input calculations have been explained clearly in the previous sections.

In feature extraction process, wavelets played major role for unsupervised method to cluster the patterns. The result of this method compared with the results of supervised method in order to cross check with the same features on.

There are some more works can be done to make the work more efficient. In the feature calculation one can try to calculate the advanced versions of daubachies transformation and can try with some other statistical approaches and apply the techniques. Also SOM and ART can be used to cluster the disease stages by making some modification in these algorithms to make the clusters as closer as possible.

References

[1]C, L., & P, B. (2004). Journal of neuroscience methods. *Online prediction of self-paced hand-movements from subthalamic activity using neural networks in Parkinsons disease* , 193-205.

[2]GA, I., & RA, N. (2008). International Journal of Psychophysiology. *Recognition of ongoing mental activity with artificial neural network* , 180.

[3]I, C. (1998). Connectionist Models in Practice and Research. *Neural network analysis of learning in autism. Neural Networks and Psychopathology* , 274.

[4]J, S. (2003). *A tutorial on Principal Component Analysis.*

[5]J, w. (2010). *Decision support for treatment of patients with advanced parkinson's disease.* Uppsala: uppsala universitet.

[6]J, W., & Ghiamati S, M. M. (2010). Journal of Neuroscience Methods. *A new computer method for assessing drawing impairment in Parkinson's disease* , 143-148.

[7]J, W., & M, D. (2010). *Analysis of Tapping Test Results in a Test Battery for Advanced PD Patients.* Uppsala: Uppsala University.

[8]J, w., M, d., D, N., & T, G. (2009). computer methods and programs in biomedicine. *A home environment test battery for status assessment in patients with advanced Parkinson's disease* , 1-4.

[9]K, L. (2002). *Wavelets theory and applications.* Chicago.

[10]M, A. (2009). *A web enabled fuzzy rule-based decision support system for dose adjustments of Duodopa infusion to patients with advanced Parkinson's disease.* Borlange.

[11]M, A. (2008). Expert Systems With Applications. *Support vector machines combined with feature selection for* , 4-7.

[12]M, M. (2008). *Web interface for a test battery for status assessments in advanced Parkinson* . Borlange: Hogskolan Dalarna .

[13]MM, H. (1992). neural clinic. *THe natural history of parkinson's disease in the prolevodopa and post-levodopa eras* , pp. 331-339.

[14]National Center for Biotechnology Information. (2010). *Pub Med Health.* Retrieved October 4, 2010, from National Center for Biotechnology Information: http://www.ncbi.nlm.nih.gov/pubmedhealth/PMH0001762/

[15]P, B., & C, B. (2005). Cognitive Systems Research. *A model of attentional impairments in autism: first steps toward a computational theory* , 193-2005.

[16]Parkinson's Disease Foundation. (2011). *Parkinson's Disease Foundation*. Retrieved February 12, 2011, from Parkinson's Disease Foundation: http://www.pdf.org/

[17]S, B., & J, M. (1998). Artificial Intelligence iin Medicine. *An artificial neural network stimulating performance of normal subjects and schizophrenics on the Wisconsin Card Sorting Test* , 123-138.

[18]Y, n. (2001). *Wavelets Made easy.* Harrisonburg: Birkhauser.