IISER KOLKATA

MA4207 Project Report

July 20, 2022

Rahul Bordoloi, Ananyapam De, Kalpana Das

Department of Mathematics and Statistics
Indian Institute of Science Education and Research, Kolkata

Contents

0.1 Abstract

We design a pattern classification technique using a Multi-Layer Perceptron for two multivariate data suffering from small sample size problem, namely the *Gastrointestinal Lesions in Regular Colonoscopy* data set and the *LSVT Voice Rehabilitation* data set.

The multi layer perceptron is a well known architecture in the field of machine learning and neural networks. It consists of input layers, hidden layers and output layers stacked together, with each node of each layer connected to each node of the next as well as the previous layer.

We also formulate a measure of *saliency* of the features, which tells us how sensitive the output is to each feature. Moreover, such saliency vectors can be used to form a network by putting a threshold on their L^2 norms and forming and edge between them. This would tell us which features cluster together and hence carry similar information about the data. Furthermore, forming a network with the help of saliency vectors allows us to study and make inferences about the original data using the summary statistics of the graph.

Chapter 1

Data Sets

1.1 Gastrointestinal Lesions in Regular Colonoscopy Data Set

This data set contains features extracted from colonoscopy videos used to detect gastrointestinal lesions. It contains 76 lesions - 15 *serrated adenomas*, 21 *hyperplastic lesions* and 40 *adenomas*. It is possible to consider this classification problem as a binary one by combining adenoma and serrated adenoma in the same class. According to this, hyperplasic lesions would belong to the class *benign* while the other two types of gastrointestinal lesions would go to the *malignant* class.

The technical goal is to maximize accuracy while minimizing false positives (lesions that do not need resection but are classified as if they do) and false negatives (lesions that do need resection but are classified as if they do not need it).

1.2 LSVT Voice Rehabilitation Data Set

This data set assess whether voice rehabilitation treatment lead to phonations considered *acceptable* or *unacceptable* (binary class classification problem). It contains 126 samples from 14 participants, 309 features.

Again, the technical goal is to maximize accuracy while minimizing false positives (voice recordings that are unacceptable but are classified acceptable) and false negatives (voice recordings that are acceptable but are classified unacceptable).

Chapter 2

Pre-processing

Data pre-processing is a data mining technique which is used to transform the raw data into a useful and efficient format. The standard steps involved in data pre-processing are

1. Cleaning - The data can have many irrelevant and missing parts. To handle this, data cleaning is done. It involves handling of missing data, noisy data etc.

2. Transformation - This step is done in order to transform the data into an appropriate form suitable for mining process. This involves transformations such as centering, normalization, discretization, attribute selection etc.

3. Reduction - When working with huge volumes of data, analysis is hard. In order to overcome this difficulty, we make use of data reduction techniques, which reduce data storage and analysis costs. Some examples are data cube aggregation, attribute subset selection, dimensionality reduction etc.

The data sets we use were already cleaned. We transform the cleaned data by centering the data to a zero mean, normalizing the data and numerically labelling the classes. The reduction techniques we employ will be discussed in the later sections.

Chapter 3

Multi-Layer Perceptron

A *multi-layer perceptron* (MLP) is a feed forward artificial neural network that generates a set of outputs from a set of inputs. A multi-layer perceptron is characterized by several layers of input nodes connected as a directed graph between the input and output layers. It uses *backpropagation* for training the neural network.

3.1 Neural Network

The neural network of multi-layer perceptron has several layers where each neuron of a layer is fully connected to each neuron of the previous and next layer. Each neuron has a certain weight associated with it. To make the model more generalizable, we add a bias neuron which has a weight associated to a fixed value of 1, which adds bias to the model. Hence, we have a network of neurons, and all that remains is to find the optimal weights and biases. The training is done using the backpropagation algorithm, where we compute the gradient of the loss and take steps opposite to it, in order to reach a global minima in the loss landscape of the model.

In our models, we use 2 *hidden layers*, with 7 and 5 neurons in the first and second hidden layer respectively. The number of output neurons depend on the particular problem. For the Gastrointestinal Lesions in Regular Colonoscopy Data Set, we use 3 neurons in the output layer and for the LSVT Voice Rehabilitation Data Set we use 2 neurons in the output.

We take a look at the flowchart of the architechture and the tabulation of the summary of the LSVT Voice Rehabilitation model for better understanding of the functioning of the model.

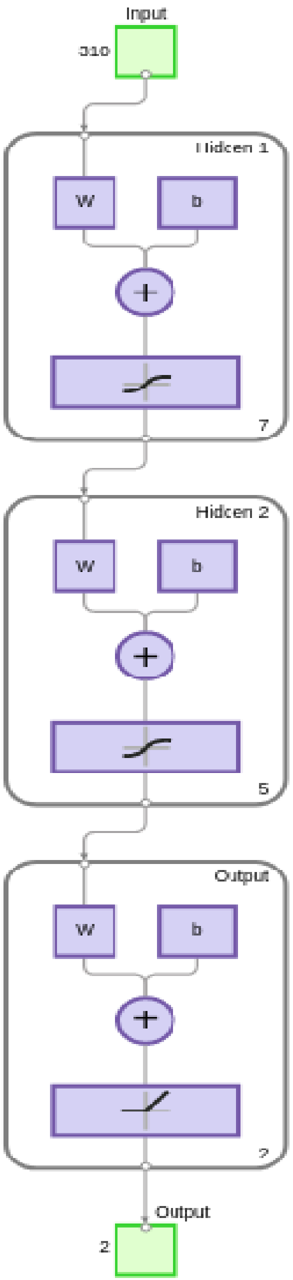
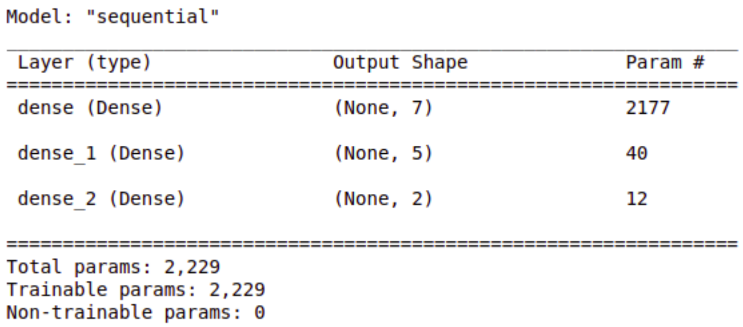


Figure 3.1: LSVT Voice Rehabilitation Model Architecture



```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 7)                 2177

 dense_1 (Dense)             (None, 5)                 40

 dense_2 (Dense)             (None, 2)                 12

=================================================================
Total params: 2,229
Trainable params: 2,229
Non-trainable params: 0
```

Figure 3.2: LSVT Voice Rehabilitation Model Summary

3.2 Activation function

In artificial neural networks, each neuron forms a weighted sum of its inputs and passes the resulting scalar value through a function referred to as an *activation function*. If a neuron has n inputs $x_1, x_2, ...x_n$, then the output of the neuron is

$$a = g(w_1x_1 + w_2x_2 + w_3x_3 + ...w_nx_n + b)$$

The function g is referred to as the activation function. The aim of activation functions is to introduce non-linearity in the system to model non-linear relationships.

In our models, we use *sigmoid activations* in all our hidden layers as well as in the output layer. The sigmoid activation function is defined as the follows

$$g(x) = \frac{1}{1 + e^{-x}}$$

where, $g(x)$ takes values in $(0, 1)$, which may also be interpreted as probabilities.

3.3 Binary Cross Entropy Loss Function

We use a *binary cross entropy loss function* as our loss function for penalizing our model predictions. This is a common categorical loss function. The formula for it is given by

$$L_y(\hat{y}) = yln(\hat{y}) + (1 - y)ln(1 - \hat{y})$$

where, $\hat{y}$ is the predicted label while y is the ground truth.

3.4 Adam Optimizer

Adaptive momentum is an algorithm for first-order gradient-based optimization of stochastic objective functions, based on adaptive estimates of lower-order moments. The method is computationally efficient, has little memory requirements,

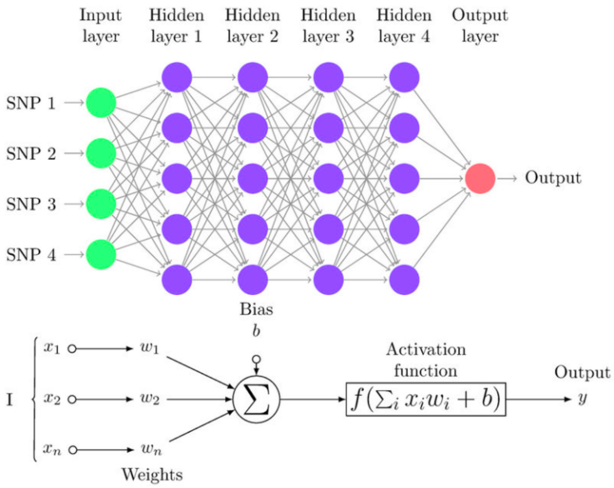


Figure 3.3: Multi-Layer Perceptron

is invariant to diagonal re-scaling of the gradients, and is well suited for problems that are large in terms of data and/ or parameters. The *hyper-parameters* have intuitive interpretations and typically require a little tuning.

The following are the equations for descent using the adam optimzer

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1)\nabla_\theta J(\theta)$$
$$v_t = \beta_2 v_{t-1} + (1 - \beta_2)\nabla_\theta^2 J(\theta)$$
$$\alpha = \eta \frac{\sqrt{1 - \beta_2^t}}{(1 - \beta_1)}$$
$$\theta_{t+1} = \theta_t - \alpha \frac{m_t}{\sqrt{v_t} + \epsilon}$$

where α is called the *learning rate* and $J(\theta)$ is the *gradient* at the point θ. Notice that the learning rate depends on the number of iterations and the gradient, hence the method is adaptive.

Chapter 4

Performance

4.1 Gastrointestinal Lesions in Regular Colonoscopy Data Set

We perform *leave-one-out cross validation* (LOOCV) for this data set even though it is generally computationally expensive because the Gastrointestinal Lesions in Regular Colonoscopy Network data set has only 152 rows, that is, it is a small data set. Moreover, this cross validation technique has almost zero randomness and very little bias.

We implement leave-one-out cross validation for the three class classification and obtain an accuracy of 86.84%, which clearly exceeds the reported accuracy of 74%. This tells that leave-one-out cross validation works very well for the three class classification with our multi-layer perceptron model.

4.2 LSVT Voice Rehabilitation Data Set

We perform *10-fold cross validation* (10-fold CV) for this data set because it is computationally less expensive compared to leave-one-out cross validation, has little bias and gives accurate estimates of the test error rate.

We implement 10-fold cross validation for the two class classification and obtain an accuracy of 79.29%, which falls within the reported accuracy range of $70-89\%$. This tells that 10-fold cross validation works very well for the two class classification with our multi-layer perceptron model.

Chapter 5

Salient Features

The sensitivity of the network's outputs to its inputs is used to rank the input features' usefulness. Firstly, an expression for the derivative of an output with respect to a given input will be derived, then it will be shown how that can be used to create a measure of the sensitivity of a trained multi-layer perceptron to each input feature. The following section will examine this measure for consistency and utility.

The notation to used throughout this section is as follows - superscripts always represent a layer index and not a quantity raised to a power. The layers are counted from the first layer of nodes which compute the sigmoid of a weighted sum. Thus, layer 1 is the first hidden layer nodes and not the inputs. The output of node i in layer j is denoted by x_i^j. Th input i is represented by x_i with no superscript, and the output i is represented by z_i. For the weights, the first subscript denotes the source node, and the second denotes the destination. The superscript on weights represents the layer of the destination node. Hence, w_{ij}^k is the weight connecting node i in layer $k-1$ to node j in layer k.

We want to calculate the derivative of the output, z_i with respect to the input, x_j. Each of the nodes in the network performs a weighted sum of its inputs plus a threshold term and puts the result through a sigmoid. Using chain rule yields,

$$\frac{\partial z_i}{\partial x_j} = z_i \left(1 - z_i\right) \frac{\partial}{\partial x_j} \left(a_i^3\right)$$

where a_i^3 is the activation of node i in layer 3. Activation is the weighted sum of the inputs plus the node threshold. Substituting the expression for the activation, we get that

$$\frac{\partial z_i}{\partial x_j} = z_i \left(1 - z_i\right) \frac{\partial}{\partial x_j} \left(\sum_m w_{mi}^3 x_m^2 + \theta_i^3\right)$$

where, w_{mi}^3 is the weight connecting node m in the second hidden layer to node i in layer $3, x_m^2$ is the output of node m in layer 2, and θ_i^3 is the threshold associated with node i in layer 3. Hence, the summation is over all nodes in layer 2. Applying the derivative to this expression for the activation gives us

$$\frac{\partial z_i}{\partial x_j} = z_i\left(1 - z_i\right)\sum_m w_{mi}^3 x_m^2\left(1 - x_m^2\right)\frac{\partial}{\partial x_j}\left(a_m^2\right)$$

where, a_m^2 is the activation of node m in layer 2. Let $\delta_i^3 = z_i\left(1 - z_i\right)$. Then,

$$\frac{\partial z_i}{\partial x_j} = \delta_i^3\sum_m w_{mi}^3 x_m^2\left(1 - x_m^2\right)\frac{\partial}{\partial x_j}\left(a_m^2\right)$$

Continuing the process through two more layers yields

$$\frac{\partial z_i}{\partial x_j} = \delta_i^3\sum_m w_{mi}^3\delta_m^2\sum_n w_{nm}^2\delta_n^1 w_{jn}^1$$

where, $\delta_m^2 = x_m^2\left(1 - x_m^2\right)$ and $\delta_n^1 = x_n^1\left(1 - x_n^1\right)$ and x_n^1 is the output of node n in layer 1. Thus, the derivative depends on the current input to the network as well as the network weights.

A measure of the saliency of an input can now be formulated as follows. Let Λ_j represent the saliency of input j, then

$$\Lambda_j = \sum_{\mathbf{x}\in\mathbf{S}}\sum_i\sum_{x_j\in D_j}\left|\frac{\partial z_i}{\partial x_j}(\mathbf{x},\mathbf{w})\right|$$

where $\mathbf{x}$ indicates the m network inputs, $\mathbf{S}$ is the set of p training vectors, w represents the weights in the network, D_j represents the set of R points for input x_j which will be sampled. Normally, D_j is a set of uniformly spaced points covering the expected range of input x_j, but in this case, as it is not computationally feasible to calculate gradients on an equally spaced 700 dimensional lattice, we calculate the gradients over the feature vectors themselves.

We then build the Gastrointestinal Lesions in Regular Colonoscopy network using the 100 most salient features from its corresponding data set, and the LSVT Voice Rehabilitation network using the 35 most salient features from its corresponding data set.

Chapter 6

Vertices and Edges

6.1 Gastrointestinal Lesions in Regular Colonoscopy Network

We construct the network for the Gastrointestinal Lesions in Regular Colonoscopy data set in the following manner

1. Let the colonoscopy images be the vertices of the network.
2. We know that the set of salient features of a colonoscopy image forms a vector. If the norm of the difference of the salient features vectors of two colonoscopy images labelled is d, then we construct an edge between the two colonoscopy images if $d < 1.3$.
3. The *threshold value* of 1.3 is obtained by selecting the median of the norms of the difference of the salient features vectors of every pair of colonoscopy images.

We visualize the graph of the Gastrointestinal Lesions in Regular Colonoscopy network, with the three different colours representing the three different classes the colonoscopy images belong to.

6.2 LSVT Voice Rehabilitation Network

We construct the network for the LSVT Voice Rehabilitation data set in a similar manner

1. Let the voice recordings be the vertices of the network.
2. We know that the set of salient features of a voice recording forms a vector. If the norm of the difference of the salient features vectors of two voice

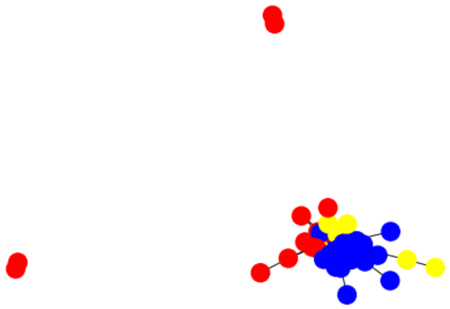


Figure 6.1: Gastrointestinal Lesions in Regular Colonoscopy Network - the three different colours represent the three different classes.

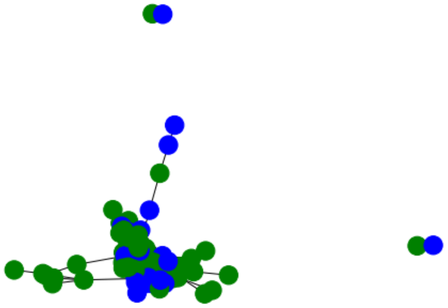


Figure 6.2: LSVT Voice Rehabilitation Network - the two different colours represent the two different classes.

recordings labelled is d, then we construct an edge between the two voice recordings if $d < 0.5$.

3. The *threshold value* of 0.5 is obtained by selecting the median of the norms of the difference of the salient features vectors of every pair of voice recordings.

Chapter 7

Network Characteristics

7.1 Gastrointestinal Lesions in Regular Colonoscopy Network

Firstly, we observe that the graph is *disconnected* with 3 *connected components*. We find the *density* of the graph to be 0.39, indicating that the graph is not close being fully connected. We find the *clustering coefficient* for each of the vertices and consequently, the *average clustering coefficient* for the graph, which turns out to be 0.71. This indicates that the graph fairly exhibits small world phenomena. This also indicates that the average of the proportion of connections of a vertex among its neighbours is fairly high. We find the *transitivity* of the graph to be 0.75, indicating that the graph's measure of triad closure is also fairly high. The results regarding the average clustering coefficient and transitivity are consistent as we expect them to be of similar values.

We plot the *giant component* of the Gastrointestinal Lesions in Regular Colonoscopy network. The vertices in the giant component are in blue and the rest of the vertices are red.

Secondly, we observe that apart from a few vertices, the entire graph is contained in the giant component. So, it wouldn't be wrong to find the *eccentricity*, *diameter* and *radius* for the giant component instead of the entire graph. These characteristics can anyway not be computed for this graph, since it is disconnected. We find the diameter to be 6 and the radius to be 3. Thus, we may conclude that the graph is similar to a *Wattz-Strogatz model* due to its real world network properties - small diameter and high average clustering coefficient.

Lastly, we compute the *centrality measures* to gauge the importance of individual vertices in a graph. We take a look at 4 different measures - *degree*, *eigenvector*, *closeness* and *betweenness*. We find that the maximum value of each of the centrality measures is 151, suggesting that the same vertex is being

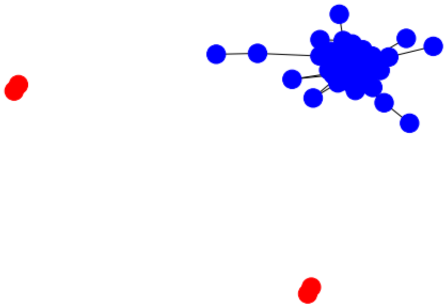


Figure 7.1: Giant Component in Gastrointestinal Lesions in Regular Colonoscopy Network

talked of in all the different cases. The maximum value of degree indicate that the most important vertex has 151 edges incident upon it. The maximum value of the eigenvector centrality measure is based not only on the number of edges incident upon it, but also based on the degree of its neighbours. The maximum value of closeness achieved in the graph indicates the minimum of the average distance from a vertex to all the other vertices. The maximum value of betweenness achieved in the graph indicates the maximum number of shortest paths a vertex has incident upon itself.

7.2 LSVT Voice Rehabilitation Network

Firstly, we observe that the graph is *disconnected* with 3 *connected components*. We find the *density* of the graph to be 0.19, indicating that the graph is far from being fully connected. We find the *clustering coefficient* for each of the vertices and consequently, the *average clustering coefficient* for the graph, which turns out to be 0.59. This indicates that the graph doesn't exhibit much of small world phenomena as the average clustering coefficient is not high. This also indicates that the average of the proportion of connections of a vertex among its neighbours is not high. We find the *transitivity* of the graph to be 0.62, indicating that the graph's measure of triad closure is also not high. The results regarding the average clustering coefficient and transitivity are consistent as we expect them to be of similar values.

We plot the *giant component* of the LSVT Voice Rehabilitation network. The vertices in the giant component are in blue and the rest of the vertices are red. We observe that all the vertices are blue because the graph is connected.

Secondly, we observe that apart from a few vertices, the entire graph is contained in the giant component. So, it wouldn't be wrong to find the *eccentricity*, *diameter* and *radius* for the giant component instead of the entire graph. These characteristics can anyway not be computed for this graph, since it is discon-

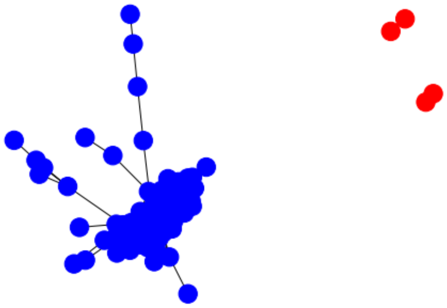


Figure 7.2: Giant Component in LSVT Voice Rehabilitation Network

nected. We find the diameter to be 9 and the radius to be 5. Thus, we may conclude that the graph is similar to a Wattz-Strogatz model due to its real world network properties - large diameter and decent average clustering coefficient. But since it doesn't have a very high average clustering coefficient, we conclude that it must be inclined towards the *regular ring lattice model.*

Lastly, we compute the *centrality measures* to gauge the importance of individual vertices in a graph. We take a look at 4 different measures - *degree*, *eigenvector*, *closeness* and *betweenness*. We find that the maximum value of each of the centrality measures is 125, suggesting that the same vertex is being talked of in all the different cases. The maximum value of degree indicate that the most important vertex has 125 edges incident upon it. The maximum value of the eigenvector centrality measure is based not only on the number of edges incident upon it, but also based on the degree of its neighbours. The maximum value of closeness achieved in the graph indicates the minimum of the average distance from a vertex to all the other vertices. The maximum value of betweenness achieved in the graph indicates the maximum number of shortest paths a vertex has incident upon itself.

Chapter 8

Clustering

8.1 Gastrointestinal Lesions in Regular Colonoscopy Network

We perform *3-mean clustering* on the salient features which we had chosen earlier. We observe that 10 iterations were required and a sum of squares errors of 135.09 was obtained. We predict the labels of the colonoscopy images based on 3-mean clustering and compare it with the ground truth, to find a match of 62.50%, which is fairly decent. We can explain the lack of a very high match due to the non linearity of the data and the fact that the data has a high variance, making it difficult to cluster data distinctly. Note that majority of the error in the match is due to the inability to distinguish serrated adenomas and adenomas clearly.

We also perform *2-mean clustering* and observe that in this case, 5 iterations were required, 147.65 sum of squares errors of was obtained and a match of 93.42% was found. We observe that the match obtained in 2-mean clustering is better than that obtained in 3-mean clustering. This can be attributed to the fact that 2 clusters is easier compared to 3 clusters, since here serrated adenomas and adenomas are considered to be in the same class.

8.2 LSVT Voice Rehabilitation Network

We perform *2-mean clustering* on the salient features which we had chosen earlier. We observe that 10 iterations were required and a sum of squares errors of 63.17 was obtained. We predict the labels of the colonoscopy images based on 2-mean clustering and compare it with the ground truth, to find a match of 54.76%, which is fairly decent. We can explain the lack of a very high match due to the non linearity of the data and the fact that the data has a high variance,

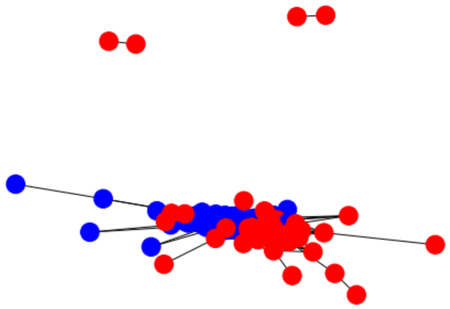


Figure 8.1: 2-mean Clustering in Gastrointestinal Lesions in Regular Colonoscopy Network

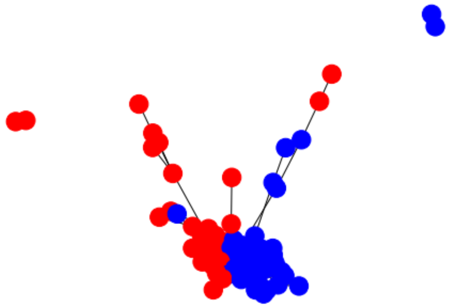


Figure 8.2: 2-mean Clustering in LSVT Voice Rehabilitation Network

making it difficult to cluster data distinctly.

Chapter 9

References

1. Feature Selection Using a Multilayer Perceptron, Dennis W. Ruck and Steven K. Rogers and Matthew Kabrisky and Wright Patterson Afb.

2. Mesejo P, Pizarro D, Abergel A, Rouquette O, Beorchia S, Poincloux L, Bartoli A. Computer-Aided Classification of Gastrointestinal Lesions in Regular Colonoscopy.

3. A. Tsanas, M.A. Little, C. Fox, L.O. Ramig: Objective automatic assessment of rehabilitative speech treatment in Parkinson's disease.

4. D. E. Rumelhart, G. E. Hinton, and R. J. Williams. 1986. Learning internal representations by error propagation. *Parallel distributed processing: explorations in the microstructure of cognition, vol. 1: foundations.* MIT Press, Cambridge, MA, USA, 318–362.