ENHANCED OUTPUT LAYER IN MULTI LAYER PERCEPTRON FOR LUNG CARCINOMA DETECTION

DR.S. KARTHIGAI
Assistant Professor
Department of Computer Science
Dr.R.A.N.M. Arts and Science College
Erode
skarthigai@yahoo.com

MR.K. GOKUL
Assistant Professor
Department of Computer Application
Dr.R.A.N.M. Arts and Science College
Erode
shreegokulbca@gmail.com

***Abstract*— Data mining techniques used in many applications as there is an incredible growth in records and it is not feasible to find a solution manually. Amongst them, the medical records has many missed values due to emergency cases or complicated situation etc. These missing values have a great influence in the desired output. The traditional mining procedure has to be enhanced to handle that missing valuesso as to get the solicited outcome. Neural network plays a vital role in many applications particularly in medical analysis. Multi Layer Perceptron is a type feed forward network has minimum three layers namely the input layer to receives the data, the hidden layer is a true computational engine, the output layer predicts the output based on the input. The perceptron train on a set of input and out pairs and learn the dependencies between them and adjust the parameters to minimize the errors. The activation function in the neuron performs the non-linear transformation function making it capable to learn and perform more complex tasks. The loss function measures the inconsistencies between the target and the predicted value. These two functions play a vital role in the whole network process. This paper focuses on these functions and made some enhancement by applying multi logit regression with Maximum A posteriori method in activation function and hyperbolic cosine function in loss function which neutralizes the learning. Also this work presents an alternative mode of training and presents a standard definition for the input parameters used in the algorithm instead of assigning numbers randomly. The proposed Enhanced Function in Multi layer Perceptron (EFMLP) is implemented in WEKA 3.9.6. and is compared with traditional MLP with suitable evaluation metrics.**

Keywords — *Multi Layer Perceptron, Multi logit regression, Maximum A Posteriori, Hyperbolic cosine loss function.*

I. INTRODUCTION

Data mining is the practice of searching large stores of data to discover hidden patterns and trends that are automated. Mining uses sophisticated mathematical algorithms to [9] partition the data and evaluate the probability of next events that may happen. It is also known as Knowledge Discovery in Data processing.

The key properties are:

- Automated discovery of patterns.
- Prediction of likely outcomes.
- Creating actionable information.
- Focusing on variant data sets.

Data are mined from anywhere whether it is stored in flat files, spreadsheets, tables or some other storage format. Thus the importance not only falls on the storage format, but its applicability over the problem to be solved.

A. Data Mining Process

The process starts from searching the data to visualizing the results in a clear and understandable format. It comprises of the below phases.

1) *Defining the problem*

This earlier phase of a mining project focuses on understanding the goals and requirements. The problem is defined, and then it is formulated and developsan implementation plan.

2) *Data Gathering and Preparation*

This phase involves data collection and exploration. Preparation phase covers all the tasks involved in creating the table used to build the model. Preparation significantly improve the mining process by formulating the problem correctly.

3) *Model Building and Evaluation*

This phase, apply various modeling techniques and fix the parameters to optimal values. If the algorithm needs data transformations while training the model, procedures are set on the basis of the method so that the final set must contain precise cases.

4) *Knowledge Deployment*

In this phase, in depth and actionable information can be derived from data. It comprise of modeling, extracting or integrating mining with the applications, making reporting tools. Mining has varied procedures for classification, grouping relation between two objects etc.

B. Neural Network

It has a collection of connected units or nodes called neurons which resembles [6] the neurons in a biological brain.
These neurons are classified across three layers:

- The input layer consists of the neurons that just receive the data and pass it on the next layer. The number of neurons in the input equals the attributes in the set.
- The output layer consists neurons and the number depends on the type of model being built.
- In between these two layers is the hidden layer, The nodes here apply transformations using functions before passing them to the upcoming layer. As the network is trained fully, the nodes that are found to

be having high prediction are weighted more heavily.

C. Multi Layer Perceptron

A Multilayer perceptron belongs to a feed forward type. It comprises of minimum three layers of nodes. Except the input, every node is a neuron that uses a nonlinear activation function. It uses a supervised learning known as back propagation for training the samples. Its multiple layers and non-linear activation differentiate it from a linear perceptron.MLP is fully connected, with a weight w_{ij} to every node with the next layer. Learning occurs by modifying the connection weight after each data is processed by having the value while the amount of error in the output is compared with the expected result. Generally, the weights are adjusted using gradient decent algorithm.

II. LITERATURE SURVEY

Akilandeswari et al [1] investigates weight optimization using Particle Swarm Optimization for Multilayer Perceptron. The features are selected by Principal Component Analysis and Hybrid PSO. Data Set Brain Computer Interface Competition is used for performance. This paper use the PSO and hybrid PSO for weight optimization in MLP. While training a neural network using PSO, the fitness value of each particle of the swarm is considered as the valueof the error function and position corresponds to the weight matrix. Mat lab is used for pre processing and WEKA tool is used for classification. From the analysis it is seen the proposed hybrid method outperforms the traditional MLP.

Bala Krishnan et al [2] designed a Intrusion detection system(IDS) based on Multi Layer Perceptron. The experiments are conducted with KddCup dataset . Four types of attacks are taken for the analysis. MLP having the feature of quick learning tendency with strong non-linear activation function. The algorithm is enhanced by selecting the features for the attacks identification randomly. The training data is pre-processed for cleaning incomplete data. The proposed IDS application with Enhanced MLP is experimented using JAVA language as front end and WEKA tool. From the result it is shown that the proposed model works effectively and it is well suitable for detecting the four types of attacks with reasonable execution time.

JaiRuby et al [3] analyse the student dataset with ID3 and MLP. The quality always resides on the dataset so it is improved in the pre processing step. The study is mainly focused on deriving a quality training dataset of educational domain. The dataset is taken from PG Computer Application course offered by an Arts and Science College between 2007 and 2012. Student personal and academic details were collected. From the result it is revealed that improved ID3 and MLP gives better result than the original dataset. Also the improved ID3 has the highest accuracy than MLP because of the speed in processing.

Mohammad et al [4] proposes Multilayer Perceptron network to predict customer churn in the leading Malaysian’s telecommunication companies. Many companies facing the loss due to the migration of the customer. This work identifies the customers that have the potential to churn in earlier stage. The results are compared against the churn prediction techniques Multiple Regression and Logistic Regression Analysis. The result has proven the supremacy over the statistical models in prediction tasks. The finding suggests that this proposed learning offer an alternative to traditional predictive approaches in customer churn prediction.

Shrwan Ram et al [5]compare Multilayer Perceptron, Radial basis and Logistic Regression to classify the Healthcare dataset. In MLP, the weighted sum of the inputs and bias are propagated to activation function to produce the output. Radial basis have Gaussian kernel functions. Logistic regression describes the relationship between the dependent binary variable and metric independent variables. Two dataset are taken for the analysis. One from the UCI and the other is collected from the local Hospital in Jodhpur. IBM SPSS and WEKA both are used for data classification and prediction. In the result it is find out the Radial basis and Logistic Regression techniques are more efficient and useful for developing the knowledge base system.

III. METHODOLOGY

A. *Existing Methodology*

Multi Layer Perceptron -

Multi layer perceptron has the [6] feature of distinguishing the data of the kind non - linear separability.
It consists of,

1) Layers

The MLP comprises of three or more layers with nonlinear activating nodes. Since they are fully connected, each node in one layer connects to every node in the following layer with a weight w_{ij}.

2) Activation Function

MLP utilize Sigmoid or Logistic function as the activation function. It is also known as a transfer function. Activation function maps the resulting values in between 0 and 1. The Non Linear activation makes the model to generalize or to adapt with variant of data and distinguish between the outputs in an easy manner.
Sigmoid or logistic Function:
It is a mathematical function have a characteristic of ‘S’ shaped curve. It is defined as bounded, differentiable and real function that all real input values has a non-negative derivative at each point. Using sigmoid the outcome is predicted as a probability since it ranges between 0 and 1.

$$O(x_i) = 1/\,(1+ e^{-xi}) \qquad \text{--- (1)}$$

Where, O is the output of the i^{th} node.
x_i is the weighted sum of the input connection.

3) Loss or cost function

The function that maps values of one or many features to a real number with cost associated with those values. The cost function in back propagation is calculated as the difference between the predicted and its actual outcome. The standard function is euclidean norm (L2 loss function).

The error is calculated by,

$$L = \sum_{j=1}^{m} (a_i - a_i^p)^2 / n \qquad \text{---(2)}$$

Where,
a_i- is the actual value,
a_i^p – predicted value.
n- the data points in the training sample.

Back Propagation –

The deep neural network finds the opt mathematical solution to turn the input to output, as it may be a linear or a non-linear relation. Back propagation is often used by the gradient descent optimization to adjust the weight of neurons with the calculation of the gradient of the loss function. This is also referred as backward propagation of errors, as the error is calculated in the output and propagated in the backward direction through the layers.

4) Mode of Learning

Learning is done by modifying the weights after each input data is processed. It is modified on the basis of the amount of error that evolved in the output when compared to the expected result and is carriedout through back propagation. One set of updating the weights for all the training sequence is called one epoch.

Mode: Stochastic Gradient Decent –
Gradient is an iterative optimization method. For finding a local minimum, gradient descent is used by taking the negative of the function in the present point. It is known as 'steepest descent'. It performs update for each training example. An alternate is gradient ascent and it takes the positive of the function to find localmaximum.
The gradient decent is calculated by the delta rule,

$$\Delta v_{ji}(m) = -\frac{\eta\, \partial E(m)}{\partial w_j(m)} x_i(m) \text{--- (3)}$$

Where, ŋ is the learning rate
x_i is the output of the previous node.
v is the new weight.
w_j is the weighted sum of the input connection.

The learning rate highly influences the speed and quality of learning. The greater value train the network fast with low accuracy and the lower value gives high accuracy but slow. So the learning rate must be applied cautiously and is applied by the user as a parameter.

Advantages

- Adaptive Learning.
- Handle non liner and complex relationship.
- Having the capability of generalization.
- Highly Fault tolerant.

Disadvantages

- Activation Function – Logistic or sigmoid cannot handle multi class or multi nominal dataset.
- Euclidean loss function – Sensitive to outliers.
- Learning rate – Fixed by the user should not have any standard definition. High rate leads to non convergence to an optimal solution while low rate leads to slow convergence to solution.
- Stochastic Gradient descent – The stochastic gradient computes gradient for all the training sample one by one at a time leads to noise and took more iteration to converge. Also several passes is taken until the algorithm converge. The sum of squared Stochastic makes frequent updates with high variance leads to fluctuation in the objective function.

B. Proposed Methodology

EFMLP (Enhanced Function in Multi Layer Perceptron)

The Proposed methodology attempt to elevate the drawbacks in the existing method. The enhanced methods replaces the traditional sigmoid activation function, alters the mode of learning, employs a new loss function instead of traditional Euclidean norm and suggests a method to define the learning rate which gives superior learning.
The proposed method consists of:
1) *Layers*
The ELMLP has three layers (Input, One Hidden and Output) with nonlinear activating nodes. They are fully connected, where each node in one layer connects to every node in the next layer with a weight wij.
2) *Activation Function*
The proposed method uses a multi nominal Logit regression[7].It can handle multi class classification task.
Multi nominal logit Function:
It is a generalized form of logistic regression that can able to handle multi [8] class classification. Multi nominal is used to predict the probabilities of variant possible outcomes of a dependent variable that are categorically distributed. For any instance there are n possible outcomes rather than two. Initially it constructs a linear predictor function.
Linear Predictor function:
Defines a score from a set of weights which are linearly combined with the explanatory features for a given instance by using a dot product as,

$$\text{Score}(X_i, m) = \alpha_m . Y_i \qquad \text{--- (4)}$$

Where,
Yi is the vector of explanatory features that describes the instance i.
αmis a vector of regression coefficient corresponds to the outcome n.
Score (Xi , n) is the score associated with assigning instance i to the category m.If there are M possible outcomes, it will run M-1 logistic regression models, in which one outcome is kept as a pivot and the rest M-1 outcomes are independently regressed against the pivot outcome.

The logit function is calculated by,

$$\Pr(X_i = M - 1) = \frac{e^{\alpha_{M-1}. Y_i}}{1 + \sum_{m=1}^{M-1} e^{\alpha_m. Y_i}} \qquad \text{--- (5)}$$

Where,

- Yi is the explanatory variable that describes observation i,
- αmis thevector of regression coefficient of the outcome k,
- M is the possible outcomes.

Estimation of regression Coefficient α:

Maximum Posteriori Estimation(MAP)–

It estimates an unknown quantity, which equals the mode value of the posterior distribution of observations. It is an extension [10] [11]of Maximum likelihood (ML) function which has a prior distribution instead of having only likelihood as the ML over the quantity.

Estimating α:

Let α be the co efficient for the observation y,

$$\alpha_{MAP} = P(\alpha|Y) = \arg_{\alpha}^{\max} \frac{P(Y|\alpha).P(\alpha)}{P(Y)} \qquad \text{--- (6)}$$

Where,

P(α|Y) is the posteriori,

P(Y| α) is the likelihood,

P(α) is the prior.

The denominator P(Y) of the posterior [12]distribution has no functional dependence with α and therefore plays no role in the optimization and ignored.

By applying logarithm,

$$\alpha_{MAP} = \arg_{\alpha}^{\max} \sum_{Y_i \in Y} \log P(Y_i|\alpha) + \log P(\alpha) \qquad \text{--- (7)}$$

Where,

log P(yi | α) is the Maximum Likelihood function.

P(α) is the conjugate prior distribution.

$$P(\alpha) = \frac{\alpha^{t-1}(1-\alpha)^{a-1}}{B(t,a)} \qquad \text{--- (8)}$$

t=1 and a=1 gives a uniform distribution.

B(t, a) is the Beta function[13] used as a normalizing constant.

$$B(t,a) = \int_0^1 b^{t-1}(1-b)^{a-1}\, dt \qquad \text{---(9)}$$

3) Loss Function

L2is normally used in the loss function. The aim of this function is to find a point that minimizes the loss to the extent. The proposed method use logarithmic hyperbolic cosine function [12] (log-cosh) function. It is the logarithm of the hyperbolic cosine of the prediction error. It is defined as a relationship between the distances of a point on a hyperbola to the origin and to the coordinate axes.

The log function is calculated as,

$$L(y, y^p) = \sum_{j=1}^{m} \mathrm{Log}(\cosh(y_i^p - y_i)) \qquad \text{--- (10)}$$

Where,

yp is the predicted value,

y is the observed value.

cosh is a hyperbolic cosine function is calculated as,

$$\cosh(y) = e^{y} + e^{-y/2} \qquad \text{---(11)}$$

4) Mode of Learning

Mini batch Gradient decent –

The traditional method has two types of learning. Stochastic and batch. This work implements a mini batch mode. A variation of the batch mode that splits the training set into small batches which is then used to calculate model error and update coefficients. The sum of the gradient over the mini-batch reduces the variance of the graduation-batch gradient balance between the robustness of stochastic gradient and the batch gradient descent. Generally Mini-batch sizes, are taken as a power of two that fits the memory requirements such as ‘32, 64, 128, 256’ and so on.

5) Training Parameters

Learning Rate –

At each iteration back propagation is used to calculate the derivative of the cost function with respect to each weight and subtract it from that weight. This will make them overcorrect and the loss will actually increase or diverge. So, each derivative is multiplied by a small value called the “learning rate” before subtract it from its corresponding weight. The rate estimates how much the current situation affects the next step. It determines how fast weights or the coefficients change.

v1 (new) = v1(old) + (learning rate)*(derivativeof cost function ofv1(old)) ---(12)

Where v1 is the associated weight.

A higher learning rate will train the network faster, possibly leads to unstable state. The lower rate trains the network very slower but leads to a stable state. So balanced learning rate should be found out. Generally, it is given at random.

The proposed method gives a standard definition for the learning rate.

Learning Rate = 1/ total of hidden neurons. ---(13)

Momentum -

The momentum term helps to prevent instabilities caused by a learning rate. When it is combined with learning rate it gives suitable and robust results by having minimum iteration cycles. Momentum estimates how much past steps will affect the next step.It is given in random as a number greater than 0. High momentum rate is always accompanied by low learning rate for optimum solution.Generally it is given at random.

The proposed method gives a standard definition for the momentum.

Momentum = 1/ total of classes --- (14)

Procedure EFMLP

Step 1: Initialize the input [ ], hidden[ ] , output layers and forward through the network to generate the output value.

Step 2: Apply the multi logit activation function by equation (5) with the training pattern to generate the output values.

Step 3: Calculate the cost or loss function by equation (11) hyperbolic cosine function.

Step 5: Training parameters learning rate and momentum is fixed by equation (13) and (14).

Step 6: Mini batch gradient decent is used to update the weights.

Step 7:The gradient of the weight is calculated by multiplying the weight's output value and input activation.

Step 8: A ratio of the gradient is subtracted from the weight to get a new weight.

Step 9: New weight is assigned and the process continues until the error is minimized.

Advantages

- Handles multi class task.
- Maximum A posterior estimation in Multi logit function minimize the probability of false negative misclassification.
- Loss function logarithmic of hyperbolic cosine function minimize the gradient quickly.
- The Learning rate and momentum controls the speed of convergence and the stability of the system.
- Mini batch Gradient decent minimizes the computational effort and requires less iterations.

Disadvantage

- High Processing time.

IV. RESULTS AND DISSCUSSION

The database is created in Microsoft excel sheet. The results are validated in WEKA 3.9.6. It expands as "Waikato Environment for Knowledge Analysis".Weka support only ARFF files.

A. *Data set*

The Lung cancer dataset are collected from a medical practitioner. It consists of 15 attributes with 3772 instances.

The fifteen attributes are Patient id, gender, chronic cough, Hemoptysis, Pain in chest, Dysponia, Cachexia, Infection in lungs, Swelling, Wheezing, Dypsnea, Clubbing in nails, Dysphasia, Tumor location and a class label with four

classes Adeno carcinoma, Squemous carcinoma, Large cell Carcinoma and Small cell Lung Carcinoma.

B. *Pre Processing*

Pre processing is an earlier stage in mining the data to clean, integrate, select and reduction of the set. In this work Gain Ratio attribute [ ] evaluation pre processing method is carried out and ten attributes are selected based on the information gain. The selected attributes are Patient id, gender, Hemoptysis, Dysponia, Cachexia, Wheezing, Dypsnea, Dysphasia, Tumor location and class label.

C. *Summary of the result*

```
=== Summary ===

Correctly Classified Instances        3740       99.1516 %
Incorrectly Classified Instances        32        0.8484 %
Kappa statistic                          0.9884
Mean absolute error                      0.0171
Root mean squared error                  0.0686
Relative absolute error                  4.6771 %
Root relative squared error             16.016  %
Total Number of Instances             3772
```

Fig. 1. Summary

Figure 1. shows the summary of the results with correctly and wrongly classified data.

D. *Confusion Matrix*

```
== Confusion Matrix ==

                          classified as
   a    b    c    d   <-- 
 583    0    8    0 |  a = adenocarcinoma
   0 1213    5    5 |  b = squemouscellcarcinoma
   0    0 1107    1 |  c = Largecellcarcinoma
   0    1   12  837 |  d = scls
```

Fig. 2. Confusion Matrix

Figure 2. show the 4x4 confusion matrix in which the entries other than the diagonal shows the wrongly classified data.

E. *Performance Evaluation*

Existing Procedure MLP is evaluated against proposed procedure Enhanced Function in Multi Layer Perceptron with nine measures.

Table 1. Evaluation measures

Evaluation Measures	MLP	EFMLP
TP Rate	0.948	0.992
FP Rate	0.024	0.008
Precision	0.954	0.994
Recall	0.948	0.992
F-Measure	0.947	0.991
MCC	0.931	0.989
ROC	0.996	1.000
PRC	0.992	0.999
Accuracy	94.8%	99.1%

Table 1. list out the performance of Adaline and MLP with nine evaluation measures. It shows the proposed EFMLP outperforms the existing Adaline.

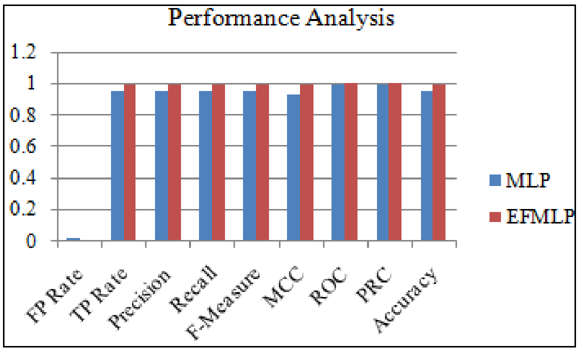

Fig. 3. Evaluation measures

Figure 3. show the various evaluation metric comparision for the existing and proposed methodology.

Figure 4. shows the acuracy level of existing and proposed methodology.

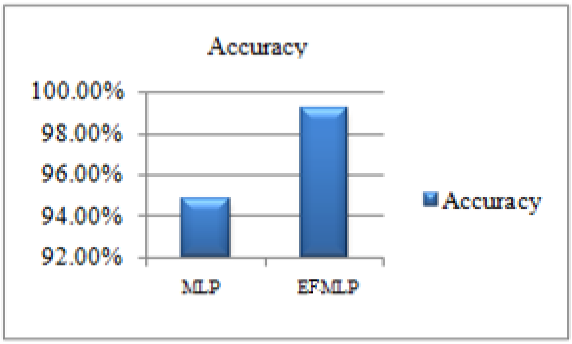

Fig. 4. Accuracy

V. CONCLUSION AND FUTURE WORK

Multi layer perceptronhas many applications notably in medical analysis. It has minimum three layers namely the input, hidden and output layer. The performance heavily relies on the training of the network. The activation and loss function have much importance in the overall accuracy. This paper focuses on these issues and enhances the activation and loss function which neutralizes the prediction. The work implements multi logit with Maximum A posteriori method and logarithmic hyperbolic cosine function as an enhancement. The enhanced method (EFMLP) is implemented in WEKA 3.9.6. and is compared with traditional MLP with suitable evaluation metrics.In future this work can be extended by enhancing the gradient decent to lessen the complexity involved.

AUTHORS PROFILE

Dr. S. Karthigai has obtained MCA Degree in Computer Applications from Shrimathi Indhra Gandhi College – Thiruchirapalli and M.phil, in Computer Science at Bharathidasan University in 2004. She has got the inclination to acquire more knowledge at present she has completed her Doctorate on Computer Science at Erode Arts and Science College - Erode. She has vast experience in teaching profession more than Thirteen years and working as an Assistant Professor and Head in computer Science at Dr.R.A.N.M. Arts and Science College – Erode. She has attended many seminars and workshops in various Universities and Colleges. She has also presented many technical papers to universities and other colleges on Computer Science and got professional papers published in many requited journals. She has acquired knowledge in Abacus and Vedic Math and has interest in teaching Mathematics. She has been awarded with certificate of merit many times for bringing out cent percent result in the university exam in Vivekananda College of Engineering for Women - Tiruchengode as well as Navarasam Arts and Science College for women - Arachalur. Since she has also published more than three books as Linux and Shell Programming and C++ Programming in Charulatha Publications.

Mr.K.Gokul has obtained M.Sc Degree in Computer Science from Erode Arts and Science College - Erode and B.Ed., in Computer Science at Holly Crass at Trichy. He has vast experience in teaching profession more than Three years and working as an Assistant Professor in computer applications at Dr.R.A.N.M. Arts and Science College – Erode. He has attended many seminars and workshops in various Universities and Colleges. He has also presented many technical papers to universities and other colleges on Computer Science and got professional papers published in many requited journals. She has acquired knowledge in Oratorical and Television Speaker in Tamil. He has been awarded with certificate of speaker in various places in Tamil Nadu. Since He has also published more than three books as C Programming in E-Book and C++ Programming in Charulatha Publications. Since He has got the inclination to acquire more knowledge at present he is doing his Doctorate on Computer Science at Erode Arts and Science College - Erode.

References

[1] Akilandeswari .K, Uma Rani .R, "Weight Optimization of Multilayer Perceptron Neural Network using Hybrid PSO for Improved Brain Computer Interface Data Classification", International Journal of Computational Intelligence and Informatics, Vol. 6: No. 4, March 2017..

[2] Bala Krishnan .R, Raajan .N.R, "An Enhanced Multilayer Perceptron Based Approach for Efficient Intrusion Detection System", International Journal of Pharmacy & Technology, 2016.

[3] Jai Ruby, David .K, "Analysis On Classification Accuracy Using Id3 And MLP with an Improved Data Set", International Journal of Pure and Applied Mathematics Volume 118 No. 18, 2018.

[4] Mohammad Ridwan Ismail, Mohd Khalid Awang, Nordin M, Rahman . A, Mokhairi Makhtar, "A Multi-Layer Perceptron Approach for Customer Churn Prediction", International Journal of Multimedia and Ubiquitous Engineering 10(7):213- 222 · July 2015..

[5] Shrwan Ram, Dr. Barwar N.C., "A Comparative Study of Multi layer Perceptron, Radial Basis Function Networks and logistic Regression for Healthcare Data Classification", International Journal of Advance ngineering and Research Development Volume 3, Issue 3, March -2016.

[6] Beale, R., & Jackson, T., "Neural Computing: An Introduction", Bristol : Hilger, c1990.

[7] Schwab, J. A. (2002). Multinomial logistic regression: Basic relationships and complete problems.

[8] Garson, G. D. (2011). "Logistic Regression", from Statnotes: Topics in Multivariate Analysis.

[9] Jiawei Han and Michelin Kamber, "Data Mining: Concepts and Techniques," Morgan Kaufmann Publishers, ISBN 1-55860-489-8. August 2000.

[10] math10.com/en/algebra/hyperbolic functions/ hyperbolic functions.html.

[11] wiseodd.github.io/techblog/2017/01/01/mle-vs-map/

[12] ragrawal.wordpress.com/2012/06/04/parameter-estimation/

[13] //en.wikipedia.org/wiki/Beta_function