ISSN: 2229-6956 (ONLINE)
DOI: 10.21917/ijsc.2020.0288

CATEGORIZATION OF LUNG CARCINOMA USING MULTILAYER PERCEPTRON IN OUTPUT LAYER

S. Karthigai and K. Meenakshi Sundaram
Department of Computer Science, Erode Arts and Science College, India

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 in data mining gains more popularity and have 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 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. This function plays a vital role in the output process. This work focus on this function and made some enhancement by applying multi logit regression with Maximum A posteriori method in activation function to handle multi-class classification The proposed Enhanced Activation function in Multi-Layer Perceptron is implemented in Weka 3.9.6 and it is compared with traditional MLP with suitable evaluation metrics.

Keywords:

Data Mining, Neural Network, Multi Layer Perceptron, Multi Logit Regression, Maximum APosteriori

1. 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 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 datasets.

Data are mined from anywhere whether it [8] 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.

Pre-processing is one of the most crucial steps in a mining process which has the concern about preparing and transforming the initial set. The medical data is often incomplete or lacking some missing values so 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 of 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 entries due to the result of the test taken and it leads in certain errors. This may be inadequate to take right decisions sometimes. To elevate this issue, preprocessing techniques gives the accurate solution. Generally the healthcare organization generates a plenty of data which are in structured, un-structured and semi structured format. The healthcare data are collected from heterogeneous sources like hospitals, clinics, doctor note, and patient records. Thereafter transforming them into a single standardized format is a must and is done by numerous existing pre-processing techniques and methods.

1.1 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.

1.1.1 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 develops an implementation plan.

1.1.2 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.

1.1.3 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.

1.1.4 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.

1.2 LUNG CARCINOMA

Lung cancer or Lung carcinoma, is a malignant tumor identified by [14] uncontrolled cell growth in the lung. This growth can penetrate beyond the lung by the process of metastasis into nearby tissue or other parts. The two main types are Small-Cell (SCLC) and Non-Small-Cell lung carcinoma (NSCLC).

Most common symptoms for both types:

- Coughing (including coughing up blood),
- Tobacco smoking,
- Weight loss,
- Shortness of breath and
- Chest pains.

1.3 NEURAL NETWORK

It has a collection of connected units or nodes called [3] neurons which resembles the neurons in a biological brain.

- These neurons are classified [12] 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, where the nodes 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.

1.4 MULTI-LAYER PERCEPTRON

A Multilayer perceptron belongs to a feed forward type. [15] 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.

1.5 PROBLEM DEFINITION

The Lung Carcinoma (LC Dataset) has four classes: Adeno carcinoma, Squamous carcinoma, Large Cell carcinoma, Small cell lung carcinoma. The output activation sigmoid function used in the traditional multi-layer perceptron gives binary output. So the existing function should be enhanced to handle multi class with high accuracy.

1.6 OBJECTIVES

- To give solution to handle multiclass.
- To increase the accuracy than the existing system.
- To make a complete Lung Carcinoma classification system.

2. LITERATURE REVIEW

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. The Brain Computer Interface Competition dataset 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 value of the error function and position corresponds to the weight matrix. Matlab is used for preprocessing 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 an 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.

Tomar et al. [4] made a survey on preprocessing and Post processing technique in data mining. The data may include several inconsistencies, missing values and irrelevant data. These all removed with the use of Pre-processing which is carried earlier. This paper elaborates the pre and post processing with various methods. Also it describes three visualization tools as it is vital in exploring the data.

Xiao et al. [5] introduce a Multiple Hidden Layers extreme Learning Machine Method. This paper proposes a multi hidden layers which obtains the characteristics of parameters from the first hidden layer. A three-hidden-layer is taken for analysis. The three hidden layer network structure has hidden neurons with the activation function. The parameters of the remaining hidden layers are obtained by introducing a new method. This method makes the actual output zero error approach as the expected hidden layer output. Depending on this, many experiments on regression and classification are done. The results shows the proposed one achieve the satisfactory results as compared with two and some other multilayer.

Panchal et al. [6] propose a hidden layer node selection method. Sales forecasting dataset is taken and the analysis is done with the metric Mean Squared Error with two different methods. First one is Back propagation and second is Conjugate gradient method. Initially the work starts with single hidden layer and two neurons and then the number of neurons is increased to train the network. From the analysis it is revealed, Back propagation method is steady but in conjugate gradient method the MSE is more fluctuate. Also the analysis found that if suitable numbers of hidden nodes are taken the better result is get with less training time. But if the number of hidden layers is increased then accuracy can be obtained to great extent but network became more complex. The obtained Mean Square Error for different experiments has been noted down and compared. Finally it is concluded that number of hidden nodes based on similarity between input data.

Panthong et al. [9] put forth a method for feature selection. This paper use of wrapper feature selection sequential forward and backward selection which is the simplest greedy search algorithm. Thirteen datasets containing variant numbers of attributes and dimensions are obtained from the UCI Machine Learning Repository. This study shows that the search technique using SFS based on the bagging algorithm using Decision Tree obtained better results in average accuracy than other methods. The benefits of the feature subset selection are an increased accuracy rate and a reduced run-time when searching multimedia data consisting of a large number of multidimensional datasets.

Vidya et al. [13] analyse the lung cancer dataset for smokers and nonsmokers using data mining classification procedure as smoking is the biggest risk factor of lung cancer and the long term smokers have the greater the risk of developing lung cancer. The paper analyse neural network, Bayesian and decision tree classifier. The duplicate data are removed in preprocessing. From the analysis it is shown Naïve Bayes algorithm outperforms the others.

3. METHODOLOGY

3.1 EXISTING MULTI LAYER PERCEPTRON

Multi-layer perceptron has the feature of distinguishing the data of the kind non-linear separability. 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}.

MLP utilize sigmoid or logistic function as the activation function [16]. 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.

3.2 EXISTING 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.

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 back propagation is the deep neural network finds the optimal [17] 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.

The advantages include adaptive learning, handle non liner and complex relationship, having the capability of generalization, and highly fault tolerant. The disadvantages include activation function - logistic or sigmoid cannot handle multi class or multi nominal dataset.

3.3 PROPOSED ENHANCED ACTIVATION FUNCTION IN MULTI-LAYER PERCEPTRON (AFMLP)

Enhanced Activation Function: The proposed method uses a multi nominal Logit regression. It can handle multi class classification task. It is a generalized form of logistic regression that can able to handle multi [11] class classification. Multi nominal logit regression 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: It 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. The score associated with assigning instance *i* to the category *n*. [7]. If there are *N* possible outcomes, it will run *N*-1 logistic regression models, in which one outcome is kept as a pivot and the rest *N*-1 outcomes are independently regressed against the pivot outcome.

3.4 ESTIMATION OF REGRESSION COEFFICIENT AND MAXIMUM POSTERIORI ESTIMATION

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

3.4.1 Procedure of AFMLP:

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 with the training pattern to generate the output values.

Step 3: Calculate the cost or loss function by equation.

Step 4: Gradient decent is used to update the weights by back propagation.

Step 5: The gradient of the weight is calculated by multiplying the weight's output value and input activation.

Step 6: A ratio of the gradient is subtracted from the weight to get a new weight.

Step 7: New weight is assigned and the process continues until the error is minimized.

The advantages include handles multi class task, and maximum aposterior estimation in multi logit function minimize the probability of false negative misclassification. The disadvantages include high processing time and use only binary class.

4. RESULTS AND DISCUSSION

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.

4.1 DATASET

The Lung cancer dataset are collected from a medical practitioner. It consists of 15 attributes with 3772 instances. The 15 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, squamous carcinoma, large cell carcinoma and small cell lung carcinoma.

4.2 PREPROCESSING

The preprocessing is an earlier stage in mining the data to clean, integrate, select and reduction of the set. In this work Gain Ratio attribute evaluation preprocessing method is carried out and ten attributes are selected based on the information gain.

4.2.1 Gain Ratio Attributes Evaluation:

The Gain Ratio estimate the ratio of the method Information Gain (IG). Information gain assesses the importance a given attribute. Information gain ratio is a ratio of information gain to the intrinsic value Information gain is the measure of information gained by knowing the value of the attribute, which is calculated by the entropy of the distribution before the split minus the entropy after the split. The largest information gain is equals the smallest entropy. The entropy is the average rate at which information is produced by a hypothetical source. The measure of information entropy is the negative logarithm of the probability mass function. Thus, when the data source has a lower probability value, the event carries more information than when the source data has a higher value. The selected attributes are: patient id, gender, hemoptysis, dysponia, cachexia, wheezing, dypsnea and dysphasia, tumor location with four trials and class label with four labels.

4.3 RESULTS

The results are evaluated in WEKA 3.9.6 and the screens are tabulated in Table.1.

Table.1. Summary

Metrics	Values	Percentage
Correctly Classified Instances	3653	96.8452%
Incorrectly Classified Instances	119	3.1548%
Kappa Statistics	0.9569	
Mean absolute error	0.0281	
Root mean squared error	0.1095	
Relative absolute error	7.6568	
Root relative squared error	25.5811	
Total Number of Instances	3772	

The Table.1 shows the summary of the results in WEKA tool. It shows the correctly the wrongly classified instances with statistical measures. The measures show the error rates in the classified data.

Table.2. Confusion Matrix

a	b	C	d
576	14	1	0
2	1207	6	8
1	30	1061	16
7	33	1	809

The Table.2 shows the confusion matrix with the classified instances as (a) adeno (b) squamous (c) large and (d) small cell carcinoma. The diagonal denotes the correctly classified data and the others are wrongly classified data.

Table.3. Evaluation measures

Evaluation Measures	MLP	EFMLP
TP Rate	0.948	0.968
FP Rate	0.024	0.013
Precision	0.954	0.969
Recall	0.948	0.968
F-Measure	0.947	0.969
MCC	0.931	0.957
ROC	0.996	0.998
PRC	0.992	0.995
Accuracy	94.8%	96.8%

The Table.3 shows the comparison results of existing and proposed methods with nine measures. The table shows the proposed outperforms the existing by giving high accuracy.

4.4 PERFORMANCE EVALUATION

The existing procedure named MLP is evaluated against proposed procedure enhanced function in Multi-Layer perceptron with 9 measures.

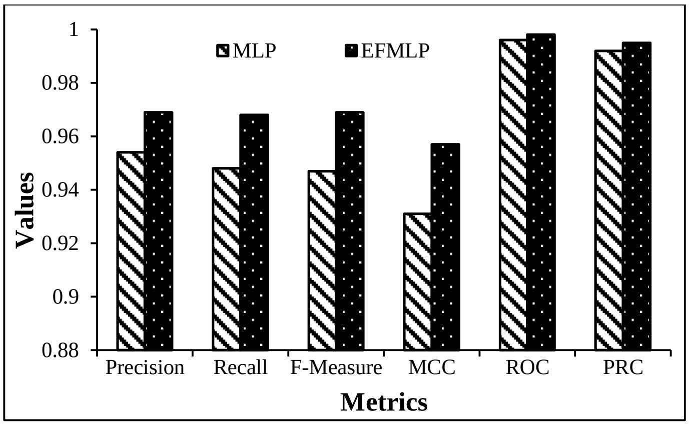

Fig.1. Evaluation measures

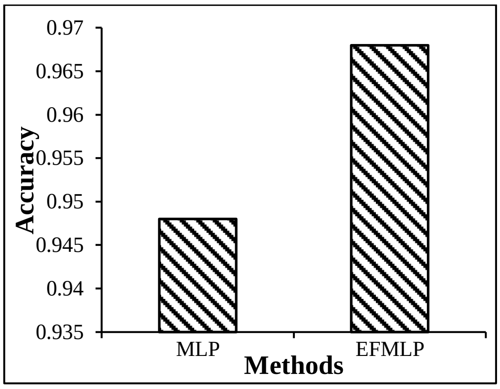

Fig.2. Accuracy

The Fig.1 shows the comparison results of existing and proposed methods. From the analysis it is found the proposed gives better result with nine measures. The Fig.2 shows the accuracy level of the existing and proposed work. The proposed method has the highest accuracy.

5. CONCLUSION AND FUTURE WORK

The Multi-layer perceptron has 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 this issue and enhances the activation function which neutralizes the prediction. The work implements multi logit with Maximum A posteriori method as an enhancement. The enhanced method (AFMLP) 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 and loss function to lessen the complexity involved.

REFERENCES

[1] K. Akilandeswari and R. Uma Rani, "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, pp. 1-12, 2017.

[2] R. Bala Krishnan .and N.R. Raajan, "An Enhanced Multilayer Perceptron Based Approach for Efficient Intrusion Detection System", *International Journal of Pharmacy and Technology*, Vol. 8, No. 4, pp. 23139-23156, 2016.

[3] R. Beale and T. Jackson, "*Neural Computing: An Introduction*", CRC Press, 1990.

[4] Divya Tomar and Sonali Agarwal, "A Survey on Pre Processing and Post Processing Technique in Mining", *International Journal of Database Theory and Applications*, Vol. 7, No. 4, pp. 21-34, 2014.

[5] Dong Xiao, Beijing Li and Yachun Mao, "A Multiple Hidden Layers Extreme Learning Machine Method and Its Application", *Mathematical Problems in Engineering*, Vol. 2017, pp. 1-10, 2017.

[6] Foram S. Panchal and Mahesh Panchal, "Review on Methods of Selecting Number of Hidden Nodes in Artificial Neural Network", *International Journal of Computer Science and Mobile Computing*, Vol. 3, No. 11, pp. 455-464, 2014.

[7] G.D. Garson, "Statnotes: Topics in Multivariate Analysis", Available at: https://faculty.chass.ncsu.edu/garson/PA765/statnote.htm.

[8] Jiawei Han and Michelin Kamber, "*Data Mining: Concepts and Techniques*", Morgan Kaufmann Publishers, 2000.

[9] Rattanawadee Panthong and Anongnart Srivihok, "Wrapper Feature Subset Selection for Dimension Reduction Based on Ensemble Learning Algorithm", *Procedia Computer Science*, Vol. 72, pp. 162-169, 2015.

[10] Sonali B. Wankhede, "Analytical Study of Neural Network Techniques: SOM, MLP and Classifier-A Survey", *IOSR Journal of Computer Engineering*, Vol. 16, No. 3, pp. 86-92, 2014.

[11] R. Vidya, V. Latha and S. Venkatesan, "Mining Lung Cancer Data for Smokers and Non-Smokers by using Data Mining Techniques", *International Journal of Trend in Research and Development*, Vol. 3, No. 7, pp. 7622-7626, 2016.