Iterative Supervised Learning Method For The Categorization Of Lung Carcinoma

[1]S.Karthigai, [2] Dr.K.Meenakshi Sundaram
[1] Research Scholar in Computer Science
Erode Arts and Science College Erode And Assistant Professor in Computer Applications,
Navarasam Arts and Science College for Women, Arachalur, India.
[2] Associate Professor of Computer Science,
Erode Arts And Science College Erode, India.

Abstract : Lung cancer in uncontrolled growth lead to tumors usually in the cell where the air pass by. They undermine the lung's ability to provide oxygen rich bloodstream. Tumors that persist in one area that doesn't spread are "benign tumors". The most dangerous one, which spread to other parts of the body are "Malignant tumors". These tumors causes early mortality. To avoid such hazardous case, prior detection with proper classification of the category of the medical condition is needed. The process of classification manually is tough. Data mining ease this process by analyzing health records from different perspectives and summarize into precise and valid information. Mining techniques has Supervised and Unsuprvised learning procedures for classification and clustering respectively. Generally Classification in supervised method categorizes and clustering in unsupervised method groups the patient records. While employing those procedures accuracy is crucial. This paper analyses the Supervised learning procedure by handling Multi layer perceptron (MLP) for the classification of lung carcimnoma. It consists of input, hidden and output layers. The accuracy is improved by the way of enhancement in the input layer by adding iterative optimization method. The proposed Iterative Multi Layer Perceptron (IMLP) promote the accuracy to a greater height than the traditional MLP. The performance is measured with nine evaluation metrics. This analysis is carried out in WEKA 3.8.6. The results are verified with variant output formats.

Keywords: Data Mining Lung Carcinoma, Iterative optimizer , MLP, IMLP.

1. Introduction

Data mining is the process of analyzing the patterns of data that are hidden naturally, according to different perspectives [7] for categorization of relevant information, which is collected and assembled in data warehouses, for effective analysis. Mining algorithm facilitates decision making and other information requirements to ultimately reduce the costs and increase the turnover. It is also known as 'Data in knowledge discovery'. Apart from analysis, it involves process such as data pre-processing and data management. It also consider model and inference , metrics, complexity , post-processing of discovered patterns, visualization and online updating. The other terms that resemble the mining are 'Data dredging, Data fishing and Data snooping'.

Stages in Mining –

1. Analysis and Selection -

The decision makers need to formulate goals, problem and objectives must be clearly defined. The process could not be proceed without the idea of the outcome. Selection includes finding the best source databases for the requirement.

2. Pre processing -

While creating the data store or warehouse, it is integrated from various sources. So there is a possibility for missing data, data conflicts and data ambiguity. To avoid this circumstances data is cleaned in this stage.

3. Transformation -

Data are transformed from one format to another format, which is more appropriate for mining. Some techniques are smoothing, Aggregating, Normalization etc.

4. Data Mining -

The sample of data is put against relevant techniques in data mining. Classification , clustering or Association rule mining are applied until a suitable methods is selected for further exploration, testing and validation.

5. Interpretation/evaluation -

Explaining the results to the decision makers is an important step in the mining process. For each technique the results are evaluated and their significance is interpreted. Most data mining tools has visualization modules. These tools communicate the results with more than two dimensions.

Supervised Learning

The task of inferring a function from a labelled training set which consist of training examples. In supervised task, [8] each sample is a pair consists of an input object and a desired output also called the supervisory signal. Supervised learning is where you have input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output.

Working Principle:

Given a set of D training samples of the form $\{(x_1, y_1),..,(x_D, y_D)\}$ such that x_i is the feature of the i-th example and y_i is its label, a learning algorithm is obtained by,

$$g : X \rightarrow Y \quad \text{--- (1.1)}$$

where X is the input space and Y is the output space. The function g is an element of possible functions D, called as the hypothesis space.

Steps in Supervised Learning –

1. Analyse the type of training examples.
2. Collect the training set.
3. Fix the input feature representation of the learned function.
4. Determine the corresponding learning algorithm.
5. Run the learning algorithm on the gathered training set.
6. Evaluate the accuracy of the learned function. The performance should be measured on a test set that is separate from the training set.

Factors to consider in Supervised Learning -

a. Heterogeneity of the data.
b. Redundancy in the data.
c. Presence of interactions and non-linearities.

Lung Carcinoma

Lung cancer or carcinoma is the number one cause of deaths in both men and women in worldwide. Smoking is the main risk factor for development of cancer. [13] It also affects the non smokers by the pollution in air. The two variants of lung cancer are small cell lung cancers (SCLC) and non-small cell lung cancers (NSCLC) which grow and propagate in different manner. Treatment involves a combination of surgery, and therapies such as chemo, immune, targeted and radiation. The prognosis is tough as it find at an advanced stage. This types is divided on the basis of the microscopic appearance of the tumor cells especially the size and the location.

Most common symptoms :

The signs related the growth of the cancer are cough, shortness of breath, pain in chest and shoulder, wheezing as it interfere in breathing.

Problem Definition

The two types of cancers (SCLC and NSCLC) increase and spread in variant ways and may have different prognosis options, so a distinction between these two types is important. Supervised learning with high accuracy is needed as it is related to loss of life. Classical algorithm can perform better if it is enhanced in a well effective manner.

Objectives

The objective of this work is:

- To enhance the traditional algorithm for high accuracy.

- To give a better distinction for the category of the lung cancer based on the symptoms.
- To implement the work in an efficient tool for viewing the results in a clear format.
- To have variant collection of test and training set.

The reaming section includes the following:
Section II describes the related works, Section III elaborates the methodology, Section IV shows the experimental results and list out the performance of the enhancement, Section V concludes the survey and enlist the future work.

2. Literature Review

Liu *et.al* [2] presented a methods for auxiliary diagnosis efficiency for lung cancer. The paper analyse Support Vector Machine, Random Forests algorithm and Fisher discriminate methods. The diagnosis confirms Support Vector and Random Forest are higher than the other, and it is thought as the optimal classification model of lung cancer. The results show that the study on diagnosis of the lung cancer by SERS on data mining is a new type of the diagnosis tool.

Ramandeep Kaur *et al* [3] did an improvisation in the Multi layer Perceptron for the classification of lung cancer. This paper analyse the nearest neighbor with MLP and put forth a MLP-NN approach that can handle noise and reduce complexity. This paper categorizes four images as bronchitis, emphysema, pleural effusion and normal. This work initially creates the group based on the nearest neighbor with the distance function and then the process of MLP continues. The result with four measures shows that the proposed MLP-NN is slightly better than the existing MLP.

Sudip Mandal *et al* [4] propose a method to classify cancer detected patients with high accuracy. This paper assist a Multi layer Feed Forward Network to detect cancer from Microarray set and UCI Data. The model is trained with Back Propagation. Two types of validations were performed with different combination of hidden layers and nodes. From the result it was found this, model classify the data with good accuracy and leads to automated medical diagnosis system for cancer detection.

Syed Moshfeq Salaken *et al* [5] put forth a new method to extract the vital features in the multi dimensional set and then apply the artificial neural network on this learned features for lung cancer classification. High dimensional set is challenging due to ineffective attributes. This paper employs a deep auto-encoder classification mechanism is to learn about the most vital features and then trains an artificial neural with these learned ones. From the results it is shown deep learned classifiers can potentially outperform the machine learning classifiers.

Yoonsuh Jung *et al* [6] proposed a cross validation method to ease model estimation and variable selection. This work presents new method to select a candidate from each hold-out fold and average the K to get the ultimate model. As per the averaging method, the variance of the estimates can be significantly reduced. This new procedure results in more stable and efficient parameter estimation than the classical procedure. In addition the asymptotic equivalence between the proposed and classical procedures in the linear regression setting is shown. This work is carried out with real data example.

3. Methodology

Flow of work

The methodology explains the existing and the proposed work.

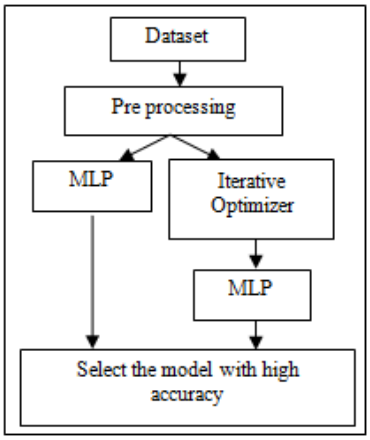


Figure 1. Flow of the proposed work

Figure 1. shows the flow of work. The existing MLP and Enhanced MLP is compared and the model with high accuracy is selected**.**

Existing method: Multi Layer Perceptron(MLP)

A multilayer perceptron (MLP) is a deep, artificial neural network. It is composed of more than one perceptron. They are composed of an input layer to receive the signal, an output layer[9] that makes a decision or prediction about the input, and in between an arbitrary number of hidden layers that are the computational engine of the MLP. MLPs with one hidden layer are capable of approximating any continuous function.

A perceptron is a linear classifier that classifies input by separating the categories with a straight line.

Input is typically a feature vector 'x' multiplied by weights 'w' and added to a bias 'b':

$$y = w * x + b \quad \text{--- (3.1)}$$

Multilayer perceptron train on a set of input pairs and learn to model .Training adjusts the parameters of the weights and biases, of the model in order to minimize error. Back propagation is used to make those weight and bias adjustments relative to the error.

Training Networks

Once configured, the neural network needs to be trained on the dataset.

A. Data Preparation

The data for training on a neural network. Data must be numerical, for example real values. If there is a categorical data, such as a sex attribute it is converted to a real-valued representation.

B. Gradient Descent

The classical [10] training algorithm for neural networks is gradient descent. This is where one row of data is given to the network at a period as input. The network processes them and finally produce an output value. This is known as forward pass. It is used in order to make predictions on new data. The output of the network is then compared to the target output and an error is calculated. This error is put back, one layer at a time, and the weights are updated according to the contributed error and this known as back propagation. The process is repeated for all of the samples in the training set.

C. Weight Updates

The weights in the network can be updated from the calculated errors and this is called online learning. It can result in fast but also make changes to the network.

Alternatively, the errors are saved up across and the network can be updated at the end. This is called batch learning and is more stable.

D. Prediction

Once a neural network has been trained it can be used to make predictions. The predictions can be made on test or validation data to estimate the performance of the model on unseen data.

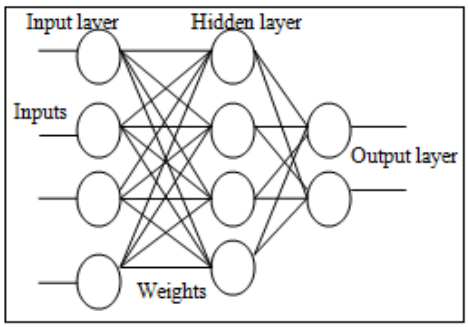


Figure 2. Architecture of MLP

Figure 2. shows the three layers in the MLP architecture with weights.

Advantages

- MLP can be used in complex applications.

Disadvantages

- The main drawback is the way it is trained.
- The number of hidden layer must be set by the user. The low value may result in under fit model and high value may result in over fit.
- Gives low accuracy in the case of missed values.
- Time Consuming.

Proposed method

IMLP – Iterative Multi layer perceptron

An approach to estimate an improve the performance of a machine learning algorithm.

[1]It works by splitting the dataset into k-parts. Each split is called a fold. Then the algorithm is trained on k-1 folds with one held back for the training and tested on the remaining each time. [11] This is repeated so that each fold is given a chance to be the held for the test. After running, k different performance scores will be get and it is summarized using a mean and a standard deviation. The result is a more accurate estimation of the performance of the algorithm on new data on each iteration. As it is trained and evaluated multiple times on different sets the result is more accurate. This is also known as rotation estimation. To lessen variability, many iterations were performed with variant partitions and then an average of the results are taken. The variance is reduced if the number of fold is increased.

Steps:

1. Randomly shuffle and partition the observations into K groups of equal length.
   { Each group is given k-1 chance for training.}
2. For the training group of observations, fit the classification model and leave the test group.
3. Use that training group's observations to test the model's predictive performance using RMSE error metrics and store this predictive information.
4. Repeat steps 2 and 3 for the next iteration.
5. Compare the target value with validated set for accuracy.
6. Select the model with high accuracy.

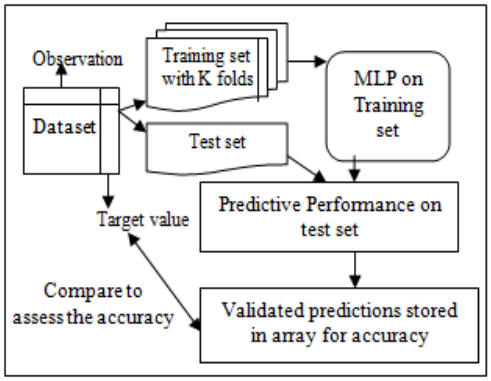


Figure 3. Steps in IMLP

Figure 3. shows the five overall steps in the proposed IMLP.

Root Mean Square Error(RMSE)

The metric used in iterative classifier optimiser is RMSE. It is calculated by [12] taking square root of the average error for all iterations and divided by the total number of iterations. It is a standard method in statistics. It represents the average error of the predicted value when compared to the actual value. The proposed method has 3 iterations.

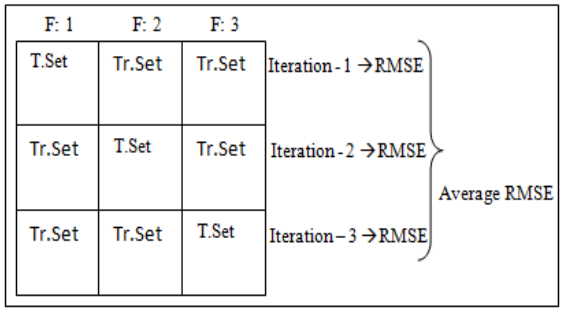


Figure 4. Iterative Optimiser

Figure 4. shows the working principle of iterative optimiser with iteration. In the figure, F is a fold number, T.Set is a test set, Tr.Set is a training set and RMSE is a Root Mean squared error.

Procedure IMLP

Step 1: Initialize the neural network with m inputs, n outputs and e connections.

Step:2 Split the set into k folds.

Step 3: Assign k-1 folds for training and the one held out for test set.

Step 4: Assign the function for the training set as,

$$y = f_n(w,x) \quad \text{--- (3.5)}$$

{The weight w maps an input x to output y)

Step 5: Phase I – Propagation

i.Propagation forward through the network to generate the output value.

ii. Calculation of the cost (error or loss term) by ,

$$E(y, y') = \tfrac{1}{2} \|y - y'\|^2 \quad \text{--- (3.6)}$$

Where y and y′ are the two outputs.

iii. Calculate the difference between the targeted and Actual output values by generating the deltas as,

$$\frac{\delta E}{\delta y'} = y' - y \quad \text{--- (3.7)}$$

Step 6: Phase II – Weight Update

i. The weight's output delta and input activation are multiplied to find the gradient of the weight.

ii. A ratio of the weight's gradient is subtracted from the weight.

Step 7: The weight must be updated in the opposite direction, "descending" the gradient.

Step 8: Record the error of validation with RMSE.

Step 9 : Change the test set by selecting the next fold.

Step 10 : Repeat from step 4 to step 7.
Step 11 : Select the model with lowest error.

Advantages

- Capable of building multiple models and then identifying the best one based on statistics.
- Evaluating multiple models from a single statement.
- Validating the robustness of the mining model.
- More accurate measure of model quality.
- All observation has the chance to get test and training set thus opposed to classical MLP training set.

Disadvantages

- The training algorithm has to be rerun k times so there is much computation.
- Iterative optimization is not supported for models that are based on the Time Series algorithm or the Sequence Clustering algorithm.

4. Experimental Results

The database is created in Microsoft excel. The results are validated in WEKA 3.8.6. It expands as "Waikato Environment for Knowledge Analysis". Weka support only Attribute Relation File Formats.

A. Data set

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

Attributes -

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 with ranker search method is carried out and ten attributes are selected based on the information gain.

Selected attributes:

Patient id, Gender, Hemoptysis, Dysponia, Cachexia, Wheezing, Dypsnea, Dysphasia, Tumor location and class label.

C. Results

Dataset in WEKA

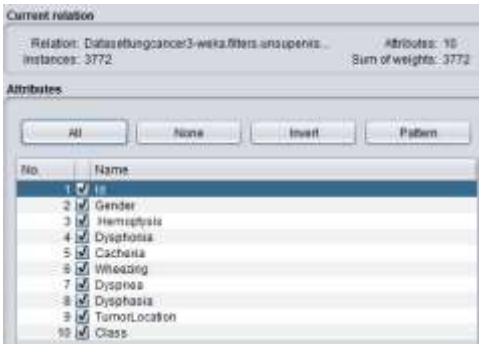
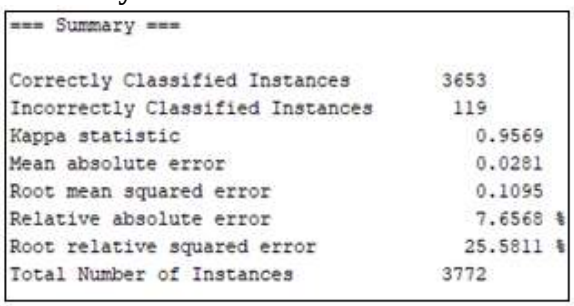


Figure 5. Loaded set in WEKA

Figure 5. shows the dataset in the tool WEKA that shows ten attributes after pre processing.

Summary of IMLP



```
=== Summary ===

Correctly Classified Instances        3653
Incorrectly Classified Instances       119
Kappa statistic                          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
```

Figure 6. Summary

Figure 6. shows the summary of the proposed IMLP result in WEKA tool.

Confusion Matrix

Table 1. Confusion matix

a	b	c	d
576	13	2	0
3	1204	5	11
1	31	1062	14
7	30	1	812

Table 1. shows the confusion matrix with 4 x 4 dimension. The entries other than in the diagonal shows the wrongly classified instances.

The data are classified into four categories as:

a- adeno carcinoma, b-squemous carcinoma,

c- largecell carcinoma d- small cell carcinoma.

D. Performance Evaluation

Table 2. MLP versus IMLP

Evaluation Measures	MLP	IMLP
TP Rate	0.948	0.968
FP Rate	0.024	0.013
Precision	0.950	0.968
Recall	0.948	0.969
F-Measure	0.947	0.967
MCC	0.931	0.957
ROC	0.996	0.998
PRC	0.992	0.995
Accuracy	94.8%	96.8%

Table 2. shows the comparison of MLP and IMLP with nine measures.

MLP versus IMLP with metrics

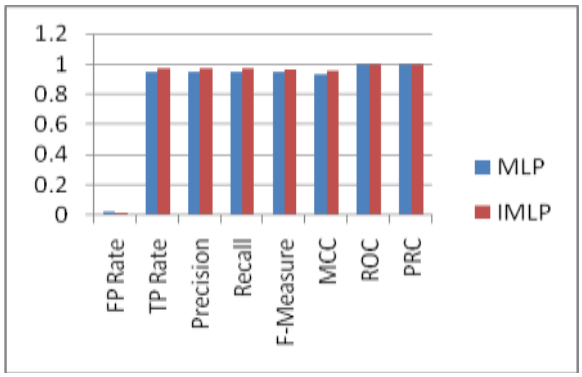


Chart 1. MLP Vs IMLP

Chart 1. Shows the performance evaluation of the existing MLP and proposed Iterative MLP with eight metrics.

Accuracy of MLP and IMLP

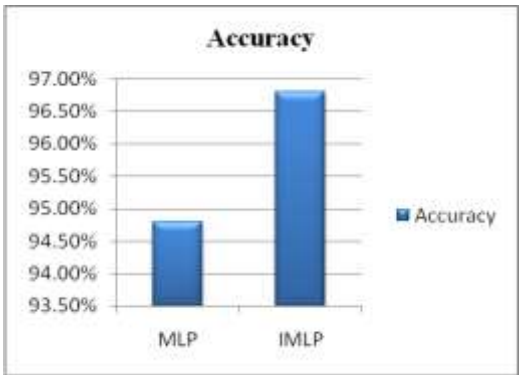


Chart 2. Accuracy

Chart 2. Shows the accuracy level of MLP and IMLP.

5. Conclusion And Future Work

Diagnosing lung cancer with large set by manual is time consuming, burden and also result in error report. Data mining techniques, ease the process by applying efficient algorithms. This work analyse the supervised learning Multi-Layer perceptron and enhance it in the input layer by adding iterative optimizer. The work is implemented in WEKA 3.8.6 with the data given by the medical practitioner. Totally fifteen attributes are taken and it is reduced in pre-processing as ten with most prominent one using procedure. From the analysis it is revealed the proposed IMLP analyse better than MLP because of the enhancement done it the input layer. It is confirmed with nine evaluation measures.

In future, the Iterative Multi-Layer Perceptron (IMLP) method can be further enhanced in the weight bias and in the hidden layer to improve the accuracy in less time.

References

[1] Keitarou Hara Ram C. Sharma, , Hidetake Hirayama, “A Machine Learning and Cross- Validation Approach for the Discrimination of Vegetation Physiognomic Types Using Satellite Based Multispectral and Multi temporal Data” Hindawi Scientifical Volume 2017 .

[2] W. Liu,"Data mining methods of Lung cancer diagnosis by saliva tests using surface enhanced Raman spectroscopy," 7th International Conference on Bio medical Engineering Infor matics, Dalian, 2014, pp. 623-627.

[3] Ramandeep Kaur, Prince Verma, “ Improved MLP-NN based approach for Lung Diseases Classification”, International Journal of Computer applications (0975 – 8887) Volume 131 – No.6, December 2015.

[4] Sudip Mandal1 and Indrojit Banerjee2, “Cancer Classification Using Neural Network” International Journal of Emerging Engineering Research and Technology Volume 3, Issue 7, July 2015.

[5] Syed Moshfeq Salaken, Abbas Khosravi, Amin Khatami, Saeid Nahavandi, Mohammad Anwar Hosen, "Lung Cancer classification Using Deep Learned Features on Low Population Dataset", IEEE 30th Canadian Conference on Electrical and Computer Engineering,2017.
[6] Yoonsuh jung, Jianhua hu, "A K-fold averaging cross-validation procedure", Journal of Non parametric, ststistics, vol 27, 2015.
[7] Arun K. Pujari, " Data mining techniques" University Press, First 2001.
[8] Daniel T.Larose, "Discovering Knowledge in Data: An Introduction to Data Mining Published by John wiley and Sons,Inc2005.
[9] en.wikipedia.org/wiki/Multilayerperceptron
[10] Machinelearningmastery.com/neural-net works-crash-course/openml.org/a/ estimation procedures.
[11] en.wikipedia.org/wiki/Crossvalidation
[12] www.developer.com/mgmt/real-world-machine-learning-model-evaluation-and-optimization.html.
[13] www. Lungcancer.wikipedia.