International Journal of Research in Engineering, Science and Management
Volume-2, Issue-5, May-2019
www.ijresm.com | ISSN (Online): 2581-5792

Diagnosis of Liver Diseases using Machine Learning

R. Yogitha[1], P. C. Manjunatha[2]
[1]*M.Tech. Student, Department of Computer Science and Engineering, Reva University, Bengaluru, India*
[2]*Assistant Professor, Department of Computer Science and Engineering, Reva University, Bengaluru, India*

***Abstract*: Liver Diseases account for over 2.4% of Indian deaths per annum. Liver disease is also difficult to diagnose in the early stages owing to subtle symptoms. Often the symptoms become apparent when it is too late. This paper aims to improve diagnosis of liver diseases by exploring 2 methods of the identification of patient parameters and genome expression. The paper also discusses the computational algorithms that can be used in the aforementioned methodology and lists demerits. It proposes methods to improve the efficiency of these algorithms.**

***Keywords*: Artificial Neural Networks, Machine Learning, SVM, Bioinformatics.**

1. Introduction

Liver disease is a tricky disease to diagnose given the subtlety of the symptoms while in the early stages. Problems with liver diseases are not discovered until it is often too late as the liver continues to function even when partially damaged Early diagnosis can potentially be life-saving. Although not discoverable to even the experienced medical practioner, the early symptoms of these diseases can be detected. Early diagnoses of patients can increase his/her life span substantially. Thus the results of this study are important both from the point of view of the computer scientist and the medical professional.

This paper aims to compare 2 methods of computer aided medical diagnoses. The first of these methods is a symptomatic approach to diagnosis. This method involves the training of an Artificial Neural Network to respond to several patient parameters such as age, Bilirubin, Alkaline Phosphatase, Alamine Aminotransferase, and Aspartate Aminotransferase among others. The Neural Network classifies the patients according to whether the patient does indeed suffer from a chronic Liver Disease or not that is healthy or not. The second method studied in this paper involves a genetic approach to the diagnosis. The proposed approach is the application of Artificial Neural Networks and Multi-Layer Perceptron to Micro-Array Analysis.

2. Architecture of liver diseases diagnosis

The following section describes the architecture of liver diseases diagnosis using machine learning:

- *Dataset:* It is a collection of related sets of information that is composed of separate elements but can be manipulated as a unit by a computer.

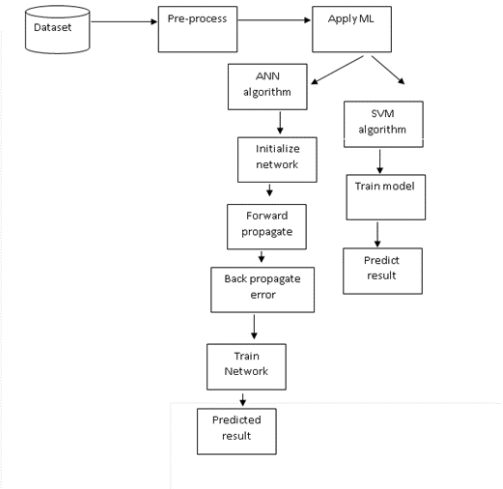

Fig. 1. Overall system architecture

- *Preprocessing:* It describes any type of processing on raw data to prepare it for another processing procedure.
- *Apply ML:* Machine Learning is the scientific study of algorithms and statistical models that computer systems use to effectively perform a specific task without using explicit instructions, relying on patterns and inference instead.
- ANN Algorithm: It is a set of algorithms that are modelled loosely after the human brain, that are designed to recognize patterns.
- *Intialize Network:* Each neuron has a set of weights that need to be maintained. The input layer is a row from our training dataset. The first real layer is the hidden layer. This is followed by the output layer that has one neuron for each class value. Initialize the network weights to small random numbers in the range of 0 to 1. Initialize network () that creates a new neural network with three input parameters, the number of inputs, the number of neurons to have in the hidden layer and the number of outputs.
- *Forward Propagate:* We can calculate an output from

a neural network by propagating an input signal through each layer until the output layer outputs its values. It is the technique we will need to generate predictions during training that will need to be corrected, and it is the method we will need after the network is trained to make predictions on new data. Forward propagation is done in three steps such as Neuron Activation, Neuron Transfer, Forward Propagation

- *Back Propagate Error:* The backpropagation algorithm is named for the way in which weights are trained. Error is calculated between the expected outputs and the outputs forward propagated from the network. These errors are then propagated backward through the network from the output layer to the hidden layer, assigning blame for the error and updating weights. This has done in two steps such as Transfer Derivative and Error Backpropagation.
- *Train Network and Predict values:* The network is trained using newly generated weights. Function named predict () is used to implements prediction. It returns the index in the network output that has the largest probability. It assumes that class values have been converted to integers starting at 0.
- SVM Algorithm: Support-vector machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis.

3. Related work

A. *Micro array analysis*

Among the most influential work in Micro-Array Analysis can be attributed to Rifkin et al [2]. Their work is attributed to a Support Vector Machine to accurately (80%) predict the origin of tumors collected from samples obtained at Massachusetts General and other medical institutions.

Kun-Hong Liu and De-Shuang Huang also solved the problems of cancer origin identification using Micro Array analysis. Several other technologies for Micro-Array analysis have been developed over the last decade. The most common ones are spotted cDNA and oligonucleotide microarrays which are discussed in this paper. Pioneers in the field include researchers from Brown and Stanford (Duggan et al Chipping Forecast 1999) where cDNA samples were hybridized to glass slides onto which the corresponding genes of interest were robotically deposited.

B. *SVM and neural networks*

Akin Ozcift and ArifGulten constructed a rotation forest ensemble classifier that was tested with success on Parkinson's, heart disease and diabetes. Some of the most useful work was done by BendiVenkataRamana et al who successfully compared various machine learning algorithms on the basis of Accuracy, Precision, Sensitivity, and Specificity when classifying this very liver patient data set. They proposed the use of Bayesian classification combined with Bagging and Boosting for improved accuracy. Bayesian classification is a simple yet powerful algorithm and works on the assumption that all variables are independent of one another. They also proposed ANOVA and MANOVA (Analysis of Variance and Multivariate Analysis of Variance) for a population comparison between the ILPD and UCI dataset.

4. Data set description

The data set used for the Neural Network Training was obtained from the online Machine Learning Repository University of California, Irvine [15]. The data was obtained from the Indian Liver Patient Data Set. The given dataset included 583 Indian Patient details. The set was first cleaned up to remove entries with missing parameters. The final set used had 583 entries, 416 of which were parameters of patients suffering from chronic Liver diseases and the remaining 167 were healthy. This data is unbalanced and thus to effectively train the classifier, we used over sampling and under sampling. The minority classes were replicated several times so as to account for a difference in the number of healthy livers versus affected livers.

5. Machine learning chemical parameters

A. *Machine learning algorithms*

1) *Back propagation*

The back-propagation algorithm is a classic multi-layered neural network algorithm developed by Rumelhart and McClelland. It works by randomizing the weights of the various layers corresponding to the input. A loss function is also defined that expresses our "unhappiness" with the result of the function. The algorithm calculates the gradient of the loss function. The parameters in the weight vectors are updated with each iteration such that they move in the direction of the absolute minimum of the loss function. The neurons are activated using ReLU (Rectified Linear Unit) or sigmoid functions.

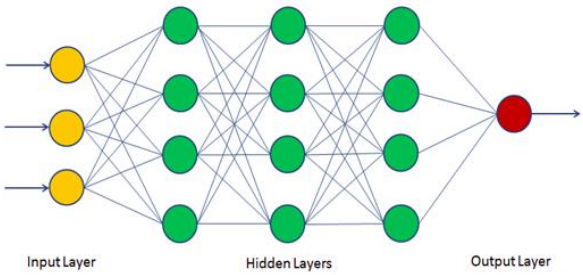

Fig. 2. Structure of Neural Network Back Propagation Algorithm

2) *Support vector machines algorithm*

A Support Vector Machine is a supervised learning algorithm. An SVM models the data into k categories, performing classification and forming an N-dimensional hyper plane. These models are very similar to neural networks. The

model was proposed by Vapnik [6]. Consider a dataset of N dimensions.

The SVM plots the training data into an N dimensioned space. The training data points are then divided into k different regions depending on their labels by hyper-planes of n different dimensions. After the testing phase is complete, the test points are plotted in the same N dimensioned plane. Depending on which region the points are located in, they are appropriately classified in that region.

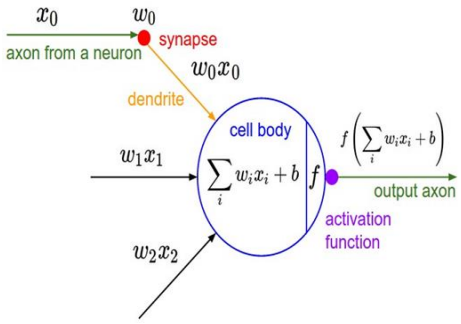

Fig. 3. Model of a neuron

B. Experimental setup

The Indian Liver Patient Dataset was obtained from Andhra Pradesh, India. The patient information included 448 male and 146 female patient records. These patients were divided into 2 groups, ones with healthy livers and ones without healthy livers.

The attributes that were considered for our experimentation were the following.

1. Age of the Patient
2. Gender of the patient
3. TB: Total Bilirubin. Bilirubin is a yellow pigment that's found in blood and stool. Excess bilirubin is a symptom of jaundice.
4. *Direct Bilirubin:* Bilirubin is of 2 types, one that is bound to a certain protein called unconjugated or indirect bilirubin. The other form, called direct bilirubin flows directly in the blood.
5. *Alkaline Phosphatase:* Alkaline Phosphatase is an enzyme that's found in the blood and helps in breaking down proteins. This is an indicator of whether the liver and gall bladder are functioning properly.
6. *Alamine Aminotransferase:* This enzyme is found in the blood and is a good indicator to verify whether a liver is damaged especially due to cirrhosis and hepatitis.
7. Aspartate Aminotransferase: Low levels of this enzyme are found in the blood. Higher level indicate damage in an organ such as heart or liver.
8. *Total Proteins:* Total proteins in the body are globulin and albumin. These levels are indicators of liver diseases.
9. *Albumin:* Albumin is the protein that prevents the fluid in blood from leaking out into the tissues.
10. *Albumin to Globulin Ratio:* It's a good indicator of the state of the liver. Normal A/G ratio is approximately 0.8 to 2.0.

C. Comparison of algorithms

The aforementioned algorithms were applied to the Indian Liver Patient Dataset (ILPD). The patient data was unbalanced in the sense that the number of affected liver patients and the number of healthy individuals were not equal. This was a difficulty during the training period. To overcome this, under-sampling and over-sampling was done. Under-sampling meant that the majority class, which in this case was the unhealthy liver set was reduced to a smaller size. Over-sampling was a method in which the minority class, in this case, the healthy individuals were replicated several times and combined with majority class.

Accuracy: Accuracy refers to the closeness of a measured value to a standard or known value.

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ and\ False\ Positive + True\ and\ False\ Negative}$$

Sensitivity: Sensitivity is also called the true positive rate measures the proportion of positives that are correctly identified.

$$Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

Precision: Precision refers to the closeness of two or more measurements to each other.

$$Precision = \frac{True\ Positives}{True\ and\ False\ Positives}$$

Specificity: Specificity is also called as true negative rate which measures the proportion of negative that are correctly identified.

$$Specificity = \frac{True\ Negative}{True\ Negatives + False\ Positives}$$

Table 1
Performance percentages of the algorithms

Algorithm	Accuracy	Precision	Sensitivity	Specificity
SVM	72	64.2	71.6	88.5
Back Propagation	73.4	65.6	73.2	87.3

6. Machine learning from microarrays

A. What is microarray analysis?

The aim of the remainder of this paper is to explain the methods of prediction of chronic liver diseases from genetic microarrays to the computer science community in an effort to direct machine learning efforts in that direction. It also proposes several tried and tested methods to conduct classification of diseased cells. The human body consists of trillions of cells which are identical for the most part in structure and shape. They carry the same number of genes and the same type of genes. However, they can be differentiated by their gene

expression when in a certain environment or when in certain conditions.

The order of information transmission in cells occurs in the following manner. The nucleus of the cell contains DNA. This DNA encodes specific information with regards to that particular cell in the form of sequences of the constituent bases, namely, adenine and thymine or guanine and cytosine. This DNA produces messenger RNA or mRNA. This mRNA then produces proteins. The complete mRNA transcript pool has been referred to as transcriptome. [8 9 10] The complete protein pool is called the proteome. For the sake of clarity, consider the muscle cells. In the nucleus of the muscle cells, DNA could be expected to generate mRNA corresponding to muscle proteins such as actin or myosin. However, it will not produce protein corresponding to the pigment melanin or the hormone insulin. Thus, the muscle cell was differentiated from the cells of the pancreas which produce insulin or from skin cells that produce melanin via the mRNA which resulted in protein production.

The above example differentiated between different types of healthy cells. A similar analysis can be carried out to differentiate between healthy cells and unhealthy cells. To differentiate between a healthy cell and a diseased one, it is possible to measure the amount of mRNA produced by every gene and compare the findings for the two cells. There are estimated to be roughly 32,000 protein encoding genes in the genome. Additionally, there are an excess of 100,000 alternately spliced transcripts from these genes. Serial Analysis of Gene Expression (SAGE) libraries help us get a better insight into the liver transcriptome. Two SAGE libraries identified nearly 15,000 to 18,000 functional transcripts related to the liver. Thus from a total of nearly 100,000 functional transcripts, 18,000 transcripts are related to the liver. The next step is to convert these mRNA transcripts into useful, mathematical quantities that can be used to predict whether a certain genome expression corresponds to a healthy liver or a diseased one. This is where microarrays are used.

B. Challenges faced

Microarray analysis of the normal human liver by Yano et al [11] shows the problems encountered when using genome expression to study the un diseased liver. 2418 genes were studied in 5 healthy patients. The study showed that only 50% of these transcripts were detected in 4 of the 5 patients. Furthermore only 27% of the gene expressions were coordinated ie. Only 27% of the genes were consistent in their expression in all 5 patients indicating the individual variability in transcript expression.

C. Developing the sample

Microarray analysis is the preferred means of determining gene expression in thousands of mRNA transcripts in a single experiment. The underlying principle of analysis remains the same although several methods have been developed after the method was first used in the early 1990s. The single strand DNA sample is applied to a substrate and the gene expression is measured versus a control DNA sample by the application of cDNA and an indicator dye that manifests the expression after hybridization occurs. The substrate may be of nylon, glass, and plastic arrays. The substrate contained grooves which contain picomoles (0.000000000001 moles) of the single stranded DNA under consideration. Thus, in the case of this experiment, 18,000 grooves corresponding to the 18,000 liver related genes required to be considered for prediction were necessary. This set up will be referred to as the microarray or a DNA chip. Two cells were necessary for the experiment, one cell each from the infected liver and a healthy liver. From the nucleus of each cell, them RNA was extracted. This mRNA was. reverse-transcript during the enzyme reverse-transcriptase which converted the mRNA into single stranded cDNA (complementary DNA). Thus samples of cDNA were obtained, one from the control or healthy cell and the other from the infected cell. These cDNA samples were then labelled using fluorescent dyes. The commonly used dyes include Cy-3 which has a wavelength of 570 nm corresponding to the green part of the visible spectrum Factors such as age, gender, ethnicity and diet continue to affect the consistency of the results and so far, have not been successfully incorporated into considered parameters.

The below figure represents the ANN's back propagation algorithm accuracy for 5 different folds. The overall accuracy arrived using Backpropagation model is around 75%. Back propagation bring more accuracy for liver disease prediction for the given datasetand Cy-5 which has a wavelength of 670 nm corresponding to the red part of the visible spectrum. For the experimental purposes, the cDNA obtained from the healthy cell was marked with Cy-3 (green) and the affected cells were marked with Cy-5 (red).

The next step was the hybridization of the cDNA with the single stranded DNA on the DNA chip. The cDNA marked with the dye and DNA on the DNA chip were mixed together using a hybridization solution, a blocking agent, and form amine. This resulted in hybridization and the cDNA strands attached onto the corresponding DNA strands. Thus 2 samples are obtained, on a single DNA chip, stained by different colored dyes, one healthy and the other affected.

To compare the two samples, the microarray was then placed into microarray scanners and the dyes manifested their fluorescence when excited by a laser of a defined wavelength. The relative intensities were then calculated which resulted in the indication of genes that expressed themselves when infected. Thus, cells from healthy and diseased livers were differentiated.

D. Methods to analyze the array

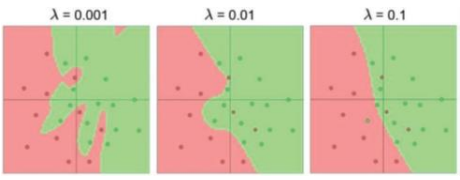

Fig. 4. Effect of regularization strength

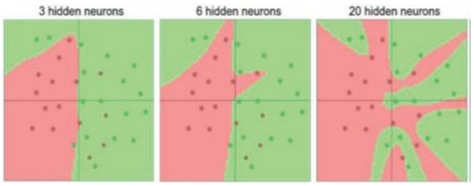

Fig. 5. Increase in predictive power with layers

E. Demerits of microarray analysis

Hepatic specific transcriptome analysis has helped us in understanding the complexities of viral hepatitis, xenobiotic metabolism and the effects of prolonged alcohol addiction and liver transplants. However, problems with this methodology continue to persist due biological noise and clinical outliers. Factors such as age, gender, ethnicity and diet continue to affect the consistency of the results and so far, have not been successfully incorporated into considered parameters.

7. Results and discussion

The below figure represents the ANN's back propagation algorithm accuracy for 5 different folds. The overall accuracy arrived using Backpropagation model is around 75%. Back propagation brings more accuracy for liver disease prediction for the given dataset.

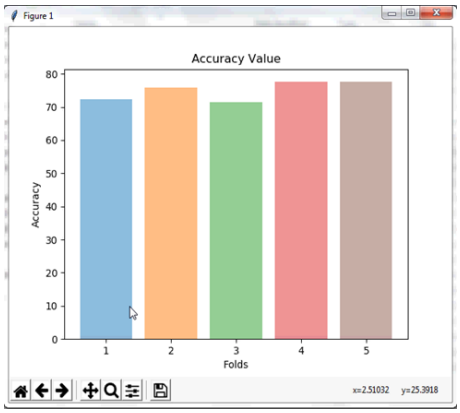

Fig. 6. Accuracy for different folds of ANN algorithm

The below figure represents the accuracy arrived for SVM classification algorithm. This algorithm arrives an accuracy of round 30%. From this study, it is visible that ANN outperforms SVM in terms of accuracy.

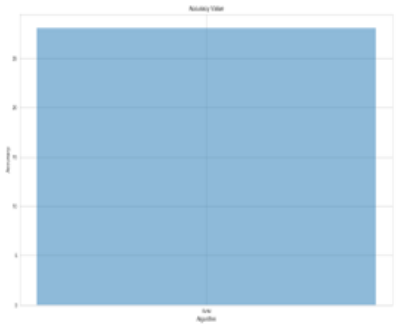

Fig. 7. Accuracy for SVM Classification algorithm

8. Conclusion

This study explores 2 methodologies in the chronic liver disease prediction such as ANN and SVM. Liver disease is especially difficult to diagnose given the subtle nature of its symptoms. Of the 2,62,6,418 deaths reported in the United States for 2014, chronic liver disease accounted for nearly 38,170 deaths. Prediction by means of computers will continue to grow in importance. In this paper 2 possibilities of machine learning models is explored that can improve the predictive power. The molecular biology approach is often affected by diet, age, and ethnicity. The chemical approach is a method of the prediction. However, in all eventuality, research in the direction of molecular biology can help to unravel the secrets to human anatomy which will help to save the lives of the people.

References

[1] Rong-Ho Lin, "An Intelligent Model for Liver Disease Diagnosis," Artificial Intelligence in Medicine, 2009.
[2] Ryan Rifkin, Sridhar Ramaswamy, Pablo Tamayo, Sayan Mukherjee, Chen-Hsiang Yeang, Micheal Angelo, Christine Ladd, Micheal Reich, Eva Latulippe, Jill P Merisov, Tomaso Poggio, William Gerald, Massimo Loda, Eric S Lander, Todd R Golub, "An Analytical Method for Multi-Class Molecular Cancer Classification ", 2003.
[3] Akin Ozcivit and Arif Gulten "Classifier Ensemble Construction with Rotation Forest to Improve Medical Diagnosis Performance of Machine Learning Algorithms", 2011.
[4] Kun-Hong Liu and De-Shuang Huang. "Cancer classification using Rotation forest", Computers in Biology and Medicine, 2008.
[5] Bendi Venkata Ramana, M. Surendra Prasad Babu and Prof. N. B. Venkateswarlu, "A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis". International Journal of Engineering Reasearch and Development, 2012.
[6] V.N. Vapnik, "Statistical Learning Theory", Wiley Publications, 1998
[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, "Delving Deep into Rectifiers", Microsoft Research, 2009.
[8] Beilharz TH, Preiss T: Translational profiling: the genome-wide measure of the nascent proteome. Brief Funct Genomic Proteomic, 2009.
[9] Gros F: From the messenger RNA saga to the transcriptome era. C R Biol. 2003, 326: 893-900.
[10] Shackel NA, Gorrell MD, McCaughan GW: Gene array analysis and the liver. Hepatology. 2002, 36: 1313-1325.