2017 International Conference on Emerging Trends & Innovation in ICT (ICEI)
Pune Institute of Computer Technology, Pune, India, Feb 3-5, 2017

Diagnosis of Liver Diseases using Machine Learning

Sumedh Sontakke[1^], Jay Lohokare[2], Reshul Dani[3*]*
[*]Department of Computer and IT, [^]Department of Electrical Engineering
College of Engineering, Pune
[1]sontakkesa15.elec@coep.ac.in, [2]lohokarejs13.comp@coep.ac.in, [3]reshulsd13.comp@coep.ac.in

***Abstract*—Liver Diseases account for over 2.4% of Indian deaths per annum. [14] 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. [1] This paper aims to improve diagnosis of liver diseases by exploring 2 methods of identification-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, Bioinformatics

I. 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 [1]. 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.

Thispaper 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 Phosphotase, 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 Perceptrons to Micro-Array Analysis .

II. 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 [4] 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[3] 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 [5] 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.

III. 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.

978-1-5090-3404-8/17/$31.00 ©2017 IEEE

IV. 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.

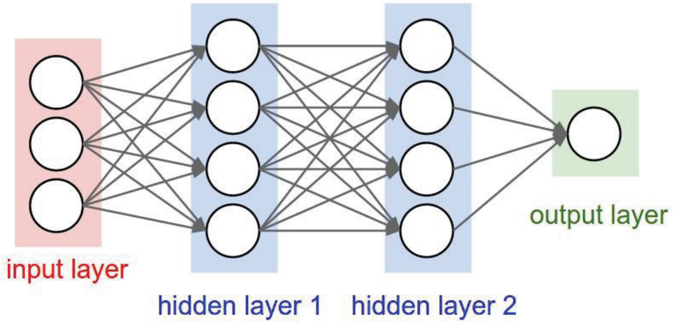

Fig 1: Structure of Neural Net Back Prop Algorithm[16]

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.

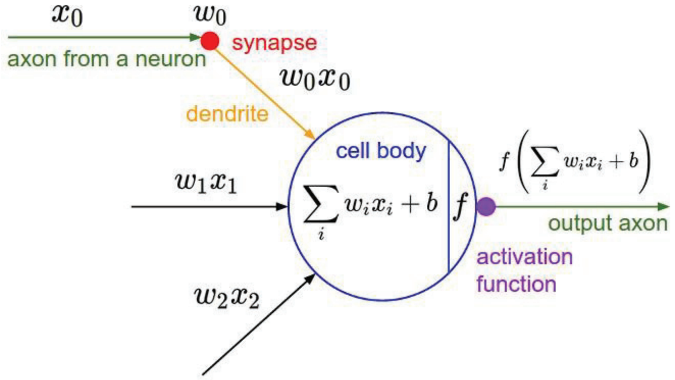

Fig 2: Model of a Neuron

B. EXPERIMENTAL SETUP

The Indian Liver Patient Dataset was obtained from Andhra Pradesh, India. The patient information included 441 male and 142 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.

i. Age of the Patient

ii. Gender of the patient

iii. TB: Total Bilirubin. Bilirubin is a yellow pigment that’s found in blood and stool. Excess bilirubin is a symptom of jaundice.

iv. 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.

v. 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.

vi. 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.

vii. Aspartate Aminotransferase: Low levels of this enzyme are found in the blood. Higher level indicate damage in an organ such as heart or liver.

viii. Total Proteins: Total proteins in the body are globulin and albumin. These levels are indicators of liver diseases.

ix. Albumin: Albumin is the protein that prevents the fluid in blood from leaking out into the tissues.

x. 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.

A. Accuracy: The accuracy of the classifier is the percentage of data points correctly classified by the algorithm.

$$\text{Accuracy} = \frac{\text{true positives + true negatives}}{\text{True and false positives + true and false negatives}}$$

B. Sensitivity: Sensitivity is the percentage of true positives correctly classified.

$$\text{Sensitivity} = \frac{\text{true positives}}{\text{False negatives + True positives}}$$

C. Precision: Precision is the percentage of true positive results versus all positive results.

$$\text{Precision} = \frac{\text{true positives}}{\text{True and false positives}}$$

D. Specificity: Specificity is the fraction of negative tuples correctly identified.

$$\text{Specificity} = \frac{\text{true negatives}}{\text{True negatives + false positives}}$$

Algorithm	Accuracy	Precision	Sensitivity	Specificity
SVM	71	64.1	71.5	88.3
Back Propagation	73.2	65.7	73.3	87.7

Table 1: Performance percentages of the algorithms

V. 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. Thus we can identify which genes express themselves and with what intensity when a change occurs in the cell, or in this case when they become diseased.

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.

Thus, microarrays are useful in allowing us to compare thousands of expressed genes in biological samples subject to various conditions. These microarrays represent 'snapshots' in time of the gene expression and are a rich source of molecular data that can be used in classification of cells and helps improve our understanding of diseased and healthy cells **.**

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 undiseased 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. Enard et al. [12] showed that samples from the same individual showed a 12% variation in gene expression and that the intraspecies variation was just as pronounced as interspecies variation in hepatic mRNA transcript expression comparing chimpanzees and humans.

Also, the complexity of transcriptome increases vastly during malignant transformations. They double and sometimes even triple in complexity.

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, the mRNAwas extracted. This mRNA was reverse-transcripted using the enzyme reverse-transcriptase which converted the mRNA into single stranded cDNA (complementary DNA). Thus two samples of cDNA were obtained, one from the control or healthy cell and the other from the infected cell. These cDNA samples were then labeled 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 and 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 formamine. 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.

Amplification is an important aspect in microarray analysis. Technologies are readily available to aid amplification [13] like Eberwine amplification, poly(A)PCR method and SMART cDNA. These methods have been successfully used to amplify the liver tissue for gene array analysis and most importantly they retain the representation of transcripts from the RNA pool.

D. METHODS TO ANALYZE THE ARRAY

The information obtained from the microarray was in the form of ratio of intensities of fluorescence from the 2 dyes. This ratio was a direct expression of the ratios of the mRNA quantities obtained from the affected cell to the mRNA quantity in the control or healthy cell. Thus, a usable mathematical form of the microarray was generated. The analysis of this form remained.

A numpy matrix of the quantities was generated. The rows corresponded to individual samples. The columns corresponded to the ratio of genes expressed. For instance, the 50th row corresponded to the 50th sample tested. Each element in this row was a ratio of the expression of a gene to the control gene expression in that column. Thus for 200 samples, the matrix obtained was 200x18000.

Linking the microarray to pathogenic occurrences can be done with or without prior knowledge of the gene function. These are called supervised and unsupervised methods of learning respectively. Another method clusters known disease states and then ascertains which disease state the unknown gene expression belongs to.

Support Vector Machines and artificial neural networks and others fall into this category.Binary classification as in this case is an intrinsically simpler problem to solve than multiclass classification. In binary classification, the algorithm can carve out the appropriate boundary for one class. The complement of this space contains the remaining data points. Ways to improve the strength of the classifier can be to reduce the regularization constant (lambda) or in the case of Artificial Neural Networks, increase the depth of the network. Fig 3 shows the improvement in predictive power after reduction in the regularization constant on a test data set for binary classification

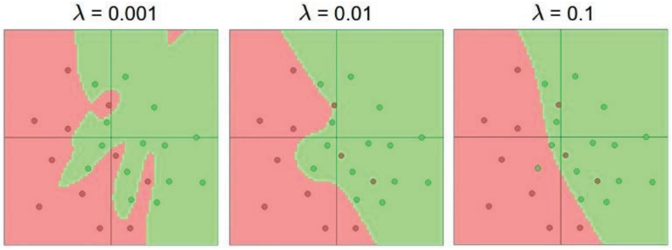

Fig 3: Effect of regularization strength[16]

However, it has been found that after 2 layers in the neural net, its ability to effectively and accurately classify test data does not increase substantially with consequent layers[16]. The strength of the classifier may also be improved by using better activation functions at neurons. Rectified Linear Unit or ReLU are experimentally found to be the better than other conventional activation functions like sigmoid and tanh.Fig 4 shows the effect of the depth of the artificial neural net on the predictive power of a test sample for binary classification.

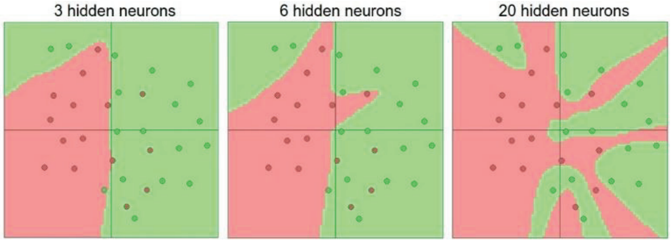

Fig 4: Increase in predictive power with layers[16]

Another problem that was faced in training binary classifier neural networks is that often neurons die in the training stage due to very high gradients passing through it. Due to this, the neuron will never fire again. This problem can be avoided by setting the 'learning rate' lower, ie. Reducing the amount by which the parameters update with every passing iteration of the loop in the direction of the global minimum of the loss function. Another solution to 'brain dead' networks is to use a "leaky" ReLU where the function has a small negative slope instead of zero. The results of the "leaky" ReLU aren't always consistent.

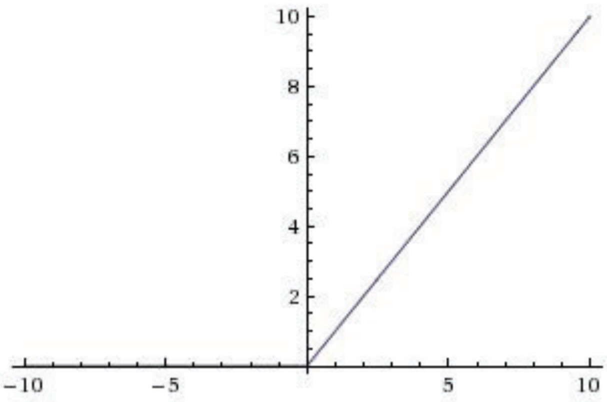

Fig 5: The Rectified Linear Unit Function (ReLU)[16]

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.

VI. CONCLUSION

The paper explores 2 methodologies in chronic liver disease prediction. Liver disease is especially difficult to diagnose given the subtle nature of its symptoms. Of the 2,626,418 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.

This paper explored 2 possibilities of machine learning models that can improve predictive power. The molecular biology approach is often affected by diet, age, and ethnicity. The chemical approach is a surer method of prediction. However in all eventuality, research in the direction of molecular biology can help unravel the secrets to human anatomy which will help save lives.

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] BendiVenkataRamana, Prof. 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. 10.1053/jhep.2002.36950.

[11] Yano N, Habib NA, Fadden KJ, Yamashita H, Mitry R, Jauregui H, Kane A, Endoh M, Rifai A: Profiling the adult human liver transcriptome: analysis by cDNA array hybridization. J Hepatol. 2001, 35: 178-186. 10.1016/S0168-8278(01)00104-0.

[12] Enard W, Khaitovich P, Klose J, Zollner S, Heissig F, Giavalisco P, Nieselt_Struwe K, Muchmore E, Varki A, Ravid R, Doxiadis GM, Bontrop RE, Paabo S: Intra- and interspecific variation in primate gene expression patterns. Science. 2002, 296: 340-343. 10.1126/science.1068996.

[13] Nicholas A Shackel, Devanshi Seth, Paul S Haber, Mark D Gorrell and Geoffrey W McCaughan, "The Hepatic Transcriptome in human Liver Disease". 10.1186/1476-5926-5-6, BioMedCentral, 2006

[14] World Health Rankings, www.worldlifeexpectancy.com

[15] UCI Machine Learning Repository http://archive.ics.uci.edu/ml/datasets/ILPD+%28Indian+Liver+Patient+Dataset%29

[16] CS231n : Convolutional Neural Networks for Visual Recognition