Classification of Liver Data using a New Algorithm

B.Saritha
MVSR Engineering college

Narra Manaswini
. Altech power and energy
systems Pvt. Ltd.

D.Hiranmayi
Altech power and energy
systems Pvt. Ltd.

S.V. Ramana
MEDCIS Pathlabs India Pvt.
Ltd.

Rama Priyanka
Altech power and energy
systems Pvt. Ltd.

K.Eswaran
Altech power and energy
systems Pvt. Ltd.
Sreenidhi Institute of Science
and Technology

ABSTRACT:
 This paper demonstrates the results of a very fast recently discovered, algorithm which was deployed to classify liver function data. This data was from a hospital in Hyderabad and the prediction of the algorithm was very accurate and it can be used for rapid initial diagnosis.

INTRODUCTION

The primary focus of this paper is to provide an example of how one can connect the rapidly expanding knowledge base of medicine, with the doctor treating each patient as an individual, with faster diagnosis by using new algorithms [1-5]. The task of ascertaining a correct diagnosis of a disease is one of the most essential problems in medicine. And Artificial Intelligence is one of the major branches of Computer Science which is most rapidly developing. AI provides several fundamental tools for intelligent data analysis.

Machine learning algorithms were designed and used to analyze medical data sets. Many researchers have used different classification algorithms to classify diseased patients from normal [6]-[13] and [15]-[17].

Today's hospitals are well equipped with data collection devices, and data is gathered and shared in large information systems. Many artificial intelligence algorithms like genetic algorithms, fuzzy logic algorithms, Bayes-methods, etc., and many data mining tools like WEKA and neural network are used to diagnose the Diseased data[6]. Some liver disease patients show symptoms of viral infection such as, tiredness, abdomen pain, muscle pain, and vomiting, and loss of appetite. But few symptoms including swelling of the abdomen as well as limbs, jaundice, digestive bleedings etc., may be the advance cases of liver failure patients. The common attributes of liver data to find the liver status are SGOT, SGPT, ALP, Total Bilirubin, Direct Bilirubin, Total Proteins and Albumin.

Algorithms and methods are applied on Medical Data in tests like LFT (Liver Function Test) to diagnose Hepatitis or liver disorder or jaundice. These patients need uninterrupted and well coordinated medical treatment to diminish the death rate.

 In this paper, a newly discovered algorithm [2, 3] which is capable of classifying and identifying liver disease data, is reported. The algorithm uses mathematical techniques to separate N number of data points, which are assumed to be d-dimensional, by a certain number of hyper-planes. The number of hyper-planes, varies from one problem to another. The best part of this algorithm is that it is non-iterative and finds the coefficients of the equations of each one of the q planes that separates all the given set of points; with a computational complexity of approximately $O(d.N.\log_2(N)) + (d^3\log_2(N))$, where N is the given number of points and d is the dimension of space. Diagnosing a liver disease is a complicated process that largely depends on the doctor's knowledge, experience, and ability to evaluate the patient's current test results and analyze risk factors that might be because of illness. In this paper we report our findings in using the Real Data Set of LFT considered which was obtained by the "MEDCIS PathLabs India Private Ltd.", Hyderabad. This database has 3750

records and 11 attributes. The proposed method has diagnosed disordered/diseased liver with an accuracy of 85.1%.

LITERATURE SURVEY

In paper [11], authors **A. Singh, B. Pandey** made a survey on combination of Artificial Intelligence methods which are used to obtain the results in classifying unhealthy liver from healthy liver. Their survey has confirmed that integrated intelligent technologies may be used for identifying liver disorders.

Bendi Venkata Ramana, Prof. M.Surendra Prasad Babu, Prof. N. B. Venkateswarlu [8], have applied Back Propagation and Support Vector Machine techniques with two data sets of which one is UCLA Liver dataset having 5 attributes and second is actual liver patient's records having 12 attributes from Andhra Pradesh.

Mehdi Neshat [9], has taken the data set from UCI machine learning repository to find out the diseased liver from healthier one. Here author considered combination of two methods namely CBRPSO and PSO. CBRPSO method was better in relation to PSO method in an average state and has higher efficiency.

Anju Gulia, Dr. Rajan Vohra, Praveen Rani, [10] have applied ILDP dataset from UCI Machine learning Repository with five different Classification algorithms. Authors in this paper showed that out of all five classification algorithms, Random Forest algorithm performance was with 72.0% accuracy which is the highest when compared to all the other techniques with the help of feature selection.

In paper [14], authors **Meherwar Fatima, Maruf Pasha**, have described and proved the use of new tools of machine learning algorithms that are used to analyze the diseases and in decision-making process accordingly.

Our paper describes the application of the Separation Algorithm on actual Medical data to classify diseased liver from healthier. The algorithm provided here separates points by planes. This can help doctor to diagnose the liver disease of the patient at the earliest, reducing the risk factor.

DATA DESCRIPTION

3750 records with eleven attributes are collected from Actual Data Set, which was provided by "Medics Path Labs India pvt. Ltd." Hyderabad.

Medical Data Details: (LFT: Liver Function Test)

Attribute Information:

1.   Bilirubin Total
2.   Bilirubin Direct
3.   Bilirubin Indirect
4.   AlkalinePhosphatase(ALP)
5.   Alanine Aminotransferase(ALT or SGPT)
6.   Aspartate Aminotransferase (AST or SGOT)
7.   Y- Glutamyl Transferase (GGT)
8.   Protein Total
9.   Albumin
10.   Globulin
11.   Albumin / Globulin Ratio

There are 2 classes: unhealthy liver and healthy liver.

Total 3750 records are divided into two sets one as training set with 3000 records and testing set with 750 records each having 11 attributes.

B.Saritha, S.V. Ramana, Narra Manaswini, Rama Priyanka, D.Hiranmayi, K.Eswaran

International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 4, Issue 9
September 2017

THE SEPARATION ALGORITHM

The algorithm [2,3], adopts an entirely new novel approach, the idea is to first treat every data point as a fundamental entity. So every data point is separated, which may be d-dimensional, from every other point by hyper planes, so that no two points are un-separated by at least one hyper plane. This seems to be counter intuitive, but the number of degrees of freedom which is always available in large d-dimensional space makes this actually not only possible but very practicable as this paper demonstrates. To give a simple example in large d-dimensional space the approximate number of hyper planes, q that are necessary to separate N points is the order of $q = \log_2(N)$ provided $N < 2^d$. This result is due to the availability of an enormous amount of degrees of freedom in large d-dimensional space, which makes it possible to orient and adjust each hyper plane in very many number of ways. Eswaran has discovered an algorithm [1-3], which does precisely this, namely given a set of N data points the algorithm discovers the q hyper planes which separate all these points from one another. This algorithm is non-iterative, it has also been demonstrated that once the data points are separated by planes, these very planes are used to classify the data points. If information is known as to which point belongs to which class is available, then one can classify a new test point by finding out its 'orientation vector'[1-3] and comparing the last with the orientation vectors of all the train points by taking dot products. We then choose that train point which has the highest dot product with the given test point and assign the class label of this train point to the test point. This procedure called Method I, classifies this test point and similarly all other test points. The additional great advantage of Method I, apart from its great speed is that the additional data can be 'added on' and the training can restart from where it last left off (there is no need to train ab-initio).This makes the algorithm exceedingly attractive because the performance will improve with time as newer and newer information is added.

There is another procedure [4] which is called "The cluster discovery algorithm", or Method II, which automatically finds the neural architecture (in a non-iterative manner) which can classify the given data. This is done by regrouping all those points which belong to the same class into d-dimensional spheres (called pure bubbles). A contiguous collection of pure bubbles, all belonging to the same class forms a cluster. Since the equations to the planes which separate each point from another is already known one can use some of these planes to separate each cluster (if needed one may have to add a few more planes) so that eventually one would have a set of planes which separate each cluster from another. The planes that separate each cluster can then form the basis of a neural architecture. Thus, in this manner one obtains a Neural architecture to solve the classification problem. All this can also be done in a non-iterative manner, (this makes the method faster than the Back-propagation algorithm), but the complexity of this Method II, is $O(N^2)$, compared to $O(N. \log_2 N)$ of Method I. See [4] for details. In this paper, we use only Method I, as it is much faster and found to be accurate also.

The detailed description along with the implementation of algorithm can be found in Refs. [1-3]. The flow chart depicts the algorithm in brief.
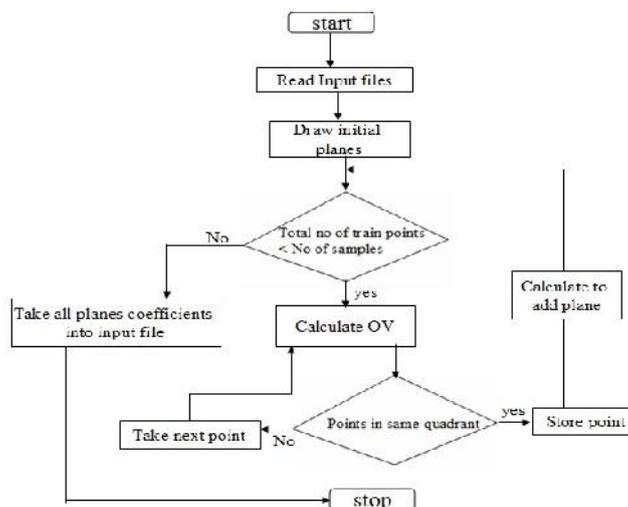

Fig 1: Flowchart "separation of points by planes".

RESULTS

The following results were obtained after applying Method I, on medical data whose values are taken from medical test called LFT's (liver function test) 11 features. We have considered 3750 unique patients LFT values. Total (3750 x 11) 41250 values are considered and the separation algorithm is applied.

) Initially number of planes considered randomly are 4 and each sample has 11 attributes (features).
) 80% of data is taken for training and 20 % of data for testing.
) Total of 40 planes were required to separate all 80 % train data.
) Time taken for training 1 second.
) We calculated the OV's (orientation vector's) for train and test points.
) Now the dot product is calculated for each test point with minimum 10% sampling from trained data.
) Now the highest dot product value and a 2 nearest neighbor is considered to find the class of the test point.
) Total time taken for the completion of testing is 1 second with 85.1% Accuracy.

CONCLUSION

Physician's responsibility is to diagnose disorders of patients and treat them at the earliest. Any wrong diagnosis can put in danger a patient's life and may even cause his/her death. In this regard, the use of different methods of artificial intelligence and expert system is a promising strategy for the future. In this paper, an algorithm called separation of points by planes has been used to classify diseased liver patients from the healthier and helped in diagnosing the dangerous hepatitis or liver disorder. This method could diagnose hepatitis or liver disorder in the best state with the accuracy of 85.1%. Total time taken for the completion of training is 1 second and testing is 1 second (on a Laptop).

REFERENCES

[1] K.Eswaran, "An Universal Classifier for Learning and Classification of data with uses in Machine Learning". Filed International Patent - Has been provisionally accepted (July 2017) by International Patent Geneva, Switzerland.

[2] K.Eswaran, "A non iterative method of separation of points by planes in n dimensions and its application". October 26, 2015, http://arxiv.org/abs/1509.08742.

[3] K.Eswaran: On non-iterative training of a neural classifier Part-I: Separation of Points by Planes. To be presented in Intellisys2017; Intelligent Systems International Conference 2017, 7-8 Sept. 2017, London, UK

[4] K.Eswaran and K.Damodhar Rao: On non-iterative training of a neural classifier Part-II: Clustering of points and their classification using an NN architecture. To be presented in Intellisys2017; Intelligent Systems International Conference 2017, 7-8 Sept. 2017, London, UK

[5] K.Eswaran, "A system and method of classification etc.", (i) Indian Patent Application No: 554/CHE/2007. Granted

(ii) Indian Patent Application No: 1256/CHE/2006 . Granted

(iii) Indian Patent Application. No. E2/1615/2016/CHE dated 24/5/2016

(Reference no. 2669/CHE/2015). [6] K.Eswaran and Vishwajeet Singh, "Some Theorems for Feed Forward Neural Networks", arXiv: 1509.05177v4 [cs.NE] 15 Oct 2015.

[6] P. J. G. Lisboa, E.C. Ifeachor and P.S.Szczepaniak, "Artificial Neural Networks in Biomedicine",Springer, London, UK, 2000.

[7] M. Neshat and M. Yaghobi, "FESHDD: fuzzy expert system for hepatitis B diseases diagnosis," in Proceedings of the 5th International Conference on Soft computing, Computing with Words and Perceptions in System Analysis, Decision and Control, (ICSCCW '09), Cyprus, September 2009.

[8] Bendi Venkata Ramana, Prof.M.Surendra Prasad Babu, Prof. N. B. Venkateswarlu, "A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis," IJDMS, Vol.3, No.2, May 2011 DOI:10.5121/ijdms.2011.3207 101.

[9] Mehdi Neshat, Mehdi Sargolzaei, Adel Nadjaran Toosi and Azra Masoumi "Hepatitis Disease Diagnosis Using Hybrid Case Based Reasoning and Particle Swarm Optimization", International Scholarly Research Network ISRN Artificia l Intelligence Volume 2012, Article ID 609718, 6 pagesdoi:10.5402/2012/609718

[10] Anju Gulia, Dr. Rajan Vohra, Praveen Rani, "Liver Patient Classification Using Intelligent Techniques", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (4), 2014, 5110-5115,

[11] Aman Singh, B. Pandey," Intelligent techniques and applications in liver disorders: a survey", Int. J. Biomed. Eng. Technol. 16(2014) 27 – 70.

[12] D. Sindhuja, R. Jemina Priyadarsini, "A Survey on Classification Techniques in Data Mining for Analyzing Liver Disease Disorder", IJCSMC, Vol.5, Issue.5, May 2016, pg.483 – 488, ISSN 2320–088X, IMPACT FACTOR: 5.258.

[13] Meherwar Fatima, Maruf Pasha, "Survey of Machine Learning Algorithms for Disease Diagnostic", Journal of Intelligent Learning Systems and Applications, 2017, 9, 1-16, http://www.scirp.org/journal/jilsa, ISSN Online: 2150-8410, ISSN Print: 2150-8402.

[14] Yogesh Singh, Pradeep Kumar Bhatia, Omprakash Sangwan, "A Review of Studies on Machine Learning Techniques", Pages -70 - 84 | Revised - 15-06-2007|Published-30-06-2007, published in IJCSS, Volume – 1 Issue-1| Publication Date - June 2007.

[15] P.Rajeswari, G.Sophia Reena,"Analysis of Liver Disorder Using Data mining Algorithm", Vol. 10 Issue 14 (Ver. 1.0) November 2010 Page | 48GJCST Classification.

[16] D. Sindhuja and R. Jemina Priyadarsini, "Liver Disease Analysis and Accuracy Prediction Using Machine Learning Techniques", I J C T A, 9(26) 2016, pp. 379-384 © International Science Press.

[17] Dr. S. Vijayarani, Mr.S.Dhayanand, "Liver Disease Prediction using SVM and Naïve Bayes Algorithms", International Journal of Science, Engineering and Technology Research (IJSETR), Volume 4, Issue 4, April 2015.