A New Approach for Liver Disease Detection using KNN and Back Propagation Algorithm

Rahul SakharamIshi, Vijay Birchha

SVCE Indore, Madhya Pradesh, India

ABSTRACT

Heavy consumption of hybrid food in today's world causes for rising of different diseases. So the study of this medical diagnosis becomes the most important part of disciplines. If there is no proper knowledge of disease then it causes serious effects. Therefore there is a requirement of strong diagnosis system. This is made possible by K-nearest neighbor algorithm and Back propagation neural network. K- Nearest algorithm is based on non-parameterized family used for regression and classification. Back propagation neural network is another technique used for diagnosis of disease based on artificial neural network used for optimization method. In this paper the comparison of both algorithms is presented and how this technique has combinable produced the better result is discussed. The combined approach provides better accuracy up to 96%. Finally frames are provided to identify disease.

Keywords: K-NN Algorithm, BPNN, Liver Disease, parameterized.

I. INTRODUCTION

Human lifestyle is changing daily with food, clothing, and shelter uses. This causes increase in medical diseases. The result of this thing is that, new disease with greater effect is arising and expert of medical field is finding difficulty to treat that patient. However, for treating this kind of diseases data mining algorithm was previously used, based on artificial intelligence[1]. The working of data mining algorithm depending on material analyzed and data used. It becomes essential technique from the physician to treat the patient. Many useful algorithms based on data mining effectively treat patient data [2]. Classification is another field of liver disease which includes prediction based approach for identification of data. Data is described with the help of vector ranging from X_1 to X_n, where n is the attribute of data. This algorithm is based on testing and training steps. Training data helped to analyzed testing data to provide solution. But the assumption is that they are sharing common distribution value.KNN Algorithm, neural network, decision tree, SVM are few of the algorithms based on the classification technique. In general it is hard to guess which algorithm is better for diagnosis of disease. KNN is one of the popular techniques used for diagnosis of disease with better performance in short period time.KNN is also having few disadvantages like importance of neighbor and its imbalanced data problem. KNN algorithm is divided into two parts, in first part need to define neighbor which are closest to each other. To get over imbalanced data problem frequency estimation and local estimation of prior probability is considered. Difference between frequency and local estimation solves the imbalanced data problem [3]. On the basis of supervised learning training method is defined and called as back propagation neural network. In this algorithm error notification is taken into consideration to improve the performance of algorithm. This algorithm consists of three layers input layer, hidden layer and output layer. By back propagating it is easy to find errors in hidden layer also. The inputs are analyzed still desired output is found. This flexibility of back propagation algorithm helps to diagnosis of disease [4]. Liver disease is one of the dangerous reasons which causes for death around the world [5]. In February 2012, approx. 2.7 to 3.9 million people are infected with hepatitis C virus in US [6].The people affected with Hepatitis B Virus 800000 to 1.4 million and 12000 people died [5][6]. In case of

India, 40% people are suffering with Hepatitis B and C and alcohol addicts are 60% [7].

Following figure 1 shows the disadvantage of alcohol consumption and disease possible due to heavy consumption of alcohol.

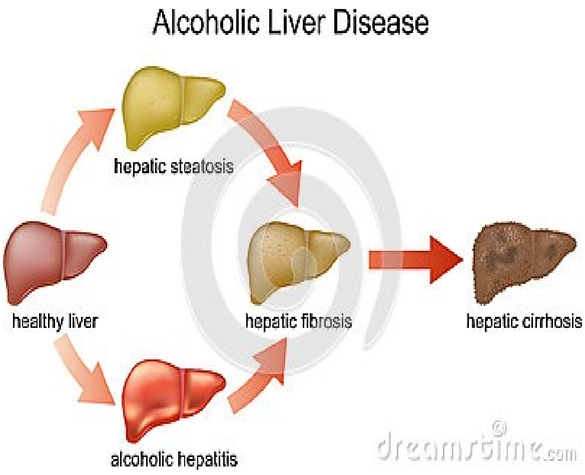


Figure 1. Disease due to alcohol Consumption

II. Literature Review

1. KNN Algorithm

This is supervised learning algorithm used for many pattern recognition, object based technique used to classify on the closet data example. The neighbor value is taken into consideration while performing classification. K is positive integer value used to search neighbor element. This algorithm consists of following steps [9]:

1. First find the number of neighbors by using k.
2. Use distance measure technique to calculate distance between instance query and training samples.
3. All training samples are sorted and kth minimum distance is found by using nearest neighbor.
4. Check sorted data and categorized the training data comes under K
5. This value of K is found with majority of nearest neighbor.

For determination of point belonging to neighbor, the distance from all training set must be calculated. Euclidean distance is technique used to get K-Nearest neighbor. For n attribute it is calculated as follow.

D= [(ft-fs1) ^2+ (ft2-fs2) ^2+…..+ (ftn - fsn) ^2] ^1/2 [1]

Smoothing parameter is needed to find next neighboring element. The value of K must be large to hold the priority ofNon-Bayesian decision, but small enough to give accurate result. The optimal value of K depends on number of sample available and structure of each population. It performs better than linear and quadratic equation to give result [10].

Advantage:

Especially if we use Inverse Square of weighted distance as the “Distance” so it becomes robust to noisy training data.If the training data is large then it is more effective.Simplicity, effectiveness, intuitiveness and competitive classification performance in many domains.

Disadvantage:

In which we need to determine value of parameter K (number of nearest neighbors).Distance based learning is not clear which type of distance to use and which attribute to use to produce the best result.

2.Back Propagation Neural Network Algorithm

It is technique based on neural network learning rule. It considered random set of objects with weight; the difference is there between neural network output and other output obtained with given input on same data set. It is used by Multi-layer Perceptron so that weights can be attached to hidden network layer to get output with back propagation algorithm. It generates error signal to change the weight value in reverse direction. The sigmoid activation is used to activate neurons while moving in forward direction [9].It uses following training algorithm to compute weight [11].

1. Initialize weight
2. repeat step 2 to 9 for stop condition
3. Input signal are provided to input unit and provided it to all hidden layer.
4. Sum of weighted input signals are provided for hidden input.
5. Output unit also performs sum of its weighted input signal and apply activation function to get computed result of output signal.
6. The target pattern is provided to output unit corresponding to input provided, error function

is computed and finally correction term is found and send it to required layer.
7. Hidden input calculates its input by considering activation function so its calculate error function and correction term with weight. It updates its weight and bias value.
8. Each output and hidden input updates its bias and weight value.
9. Stop condition is tested to get final result

Advantage:
In BPNN algorithm weight is needed to get the accuracy of data. In hidden layer the verification of data takes place. If errors are there, then it is backtracked to find the value to overcome from that error. In the data set there is backtracking to remove noise and data.

Disadvantage:
If dataset increases then backtracking algorithm time increases with the loss of efficiency. Because it needs to locate the error in that large dataset and then find the error causes for large amount of time for producing result.

III. DATA SET

The data set is of BUPA liver disorders from the source information BUPA Medical Research Ltd and donated by Richard S. Forsyth (8 Grosvenor Avenue Mapperley Park Nottingham NG3 5DX 0602-621676) on dated 5/15/1990. About Past usage none known other than what is shown in the PC/BEAGLE User's Guide (written by Richard S. Forsyth).Relevant information about data set is the first 5 variables are all blood tests which are thought to be sensitive to liver disorders that might arise from excessive alcohol consumption. Each line in the bupa data file constitutes the record of a single male individual. It appears that drinks>5 is some sort f selector on this database. See the PC/BEAGLE User's Guide for more information. Numbers of instances are 345 and number of attribute as follow:-

a) Mcv :mean corpuscular volume
b) Alkphos :alkaline phosphotase
c) Sgpt :alamine aminotransferase
d) Sgot :aspartate aminotransferase
e) Gammagt :gamma-glutamyl transpeptidase
f) Drinks number of half-pint equivalents of alcoholic beverages drunk per day
g) Selector field used to split data into two setsand also there are no missing values in data set.

IV. Proposed Approach

The performance of KNN algorithm decreases when date set is large in number. As the size of data set increases performance of algorithm also decreases [10]. It means that performance is directly proportional to dataset. In case of BPNN it uses error signal to track the fault of input but the time of producing the result of BPNN increases and accuracy decreases [11]. So to avoid the disadvantages of KNN and BPNN the proposed technique combines the advantages of KNN and BPNN algorithm to form new efficient techniqueand it avoids large computation time and improves accuracy.

KNN and back propagation both algorithms are working for diagnosis of disease. KNN considered the neighboring data set to get the diagnosis of disease and back propagation used to backtrack the data if any problem is occurred. By considering this method of backtracking, the KNN algorithm is modified in this technique and it is combined with neural network algorithm to get accurate diagnosis of liver disease to get status of malignant and benign[10]-[14]. Following are the steps of new proposed algorithm for the diagnosis of the disease.

Algorithm:

1. Find neighbors by using KNN algorithm and calculate distance using Euclidean distance method.
2. Sort the data set and find the Kth distance for data from neighbor with minimum distance.
3. Then provide the output value obtained from KNN layer to back propagation neural network method.
4. Define hidden layer and calculate weight of data obtained with kth distance obtained from KNN algorithm.
5. Use activation function to get accurate weight and bias value for data
6. If error is there the reverse the working of algorithm and corrected data with weight.
7. Then update weight and bias value for data of output layer.
8. Provide data with new value to medical expert so that the accurate diagnosis of the disease takes place.

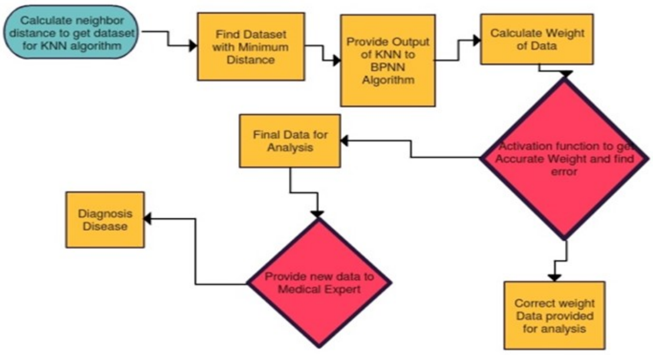

Calculate neighbor distance to get dataset for KNN algorithm
Find Dataset with Minimum Distance
Provide Output of KNN to BPNN Algorithm
Calculate Weight of Data
Activation function to get Accurate Weight and find error
Final Data for Analysis
Correct weight Data provided for analysis
Provide new data to Medical Expert
Diagnosis Disease

Figure 2. Proposed Algorithm

In this way by combining the working of KNN and BPNN algorithm the proper data is set with maximum priority is obtained to get the required data in quick time. The back tracking of BPNN helps us to identify the error associated with KNN algorithm. The data set is modified after correction of error. This accurate data set is provided for medical expert to analyze data. Then proper diagnosis is performed on that data. Result obtained with this technique requires less time for computation with maximum accuracy [15].

ADVANTAGES AND DISADVANTAGES OF CLASSIFICATION ALGORITHMS

ALGORITHMS	ADVANTAGE	DISADVANTAGE
K-Nearest Neighbor Algorithm	1. Easy to understand 2. Training is fast and robust to noisy data.	1. Limited Memory 2. Being a supervised learning lazy algorithm
Back Propagation Neural Network Algorithm	1. Capable of producing an arbitrarily complex relationship between input and output. 2. Less over fitting, robust to noise. 3. Especially popular in text classification problems	1. Do not work well when there are many hundreds or thousands of input features and difficult to understand the model. 2. SVM is a binary classifier. To do a multi-class classification, pairwise classifications can be used

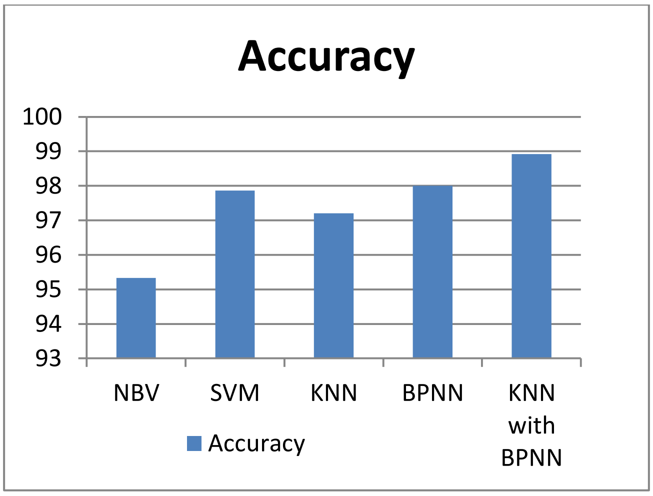
V. Result Analysis

Performance of this algorithm is measured with respect to data set consist of 583 liver patient records. The different algorithm values are shown in following table.

Sr No	Name of Technique	Training time	Testing time	Accuracy
1	KNN	0.33ms	0.15ms	97.203
2	BPNN	2.87ms	0.012ms	98.002
3	KNN with BPNN	0.28ms	0.01ms	98.92

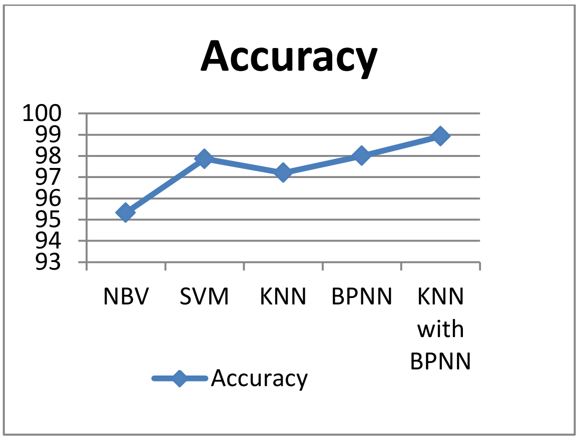
Table 1. Comparison of various algorithms

As shown in above table the proposed algorithm is tested on dataset of 583 liver patient records and we find following comparison. Naïve bayes algorithm [13] having training time and testing time more as compared to SVM algorithm [12]. If we compare SVM algorithm with KNN and BPNN algorithm, the training and testing time is more for SVM algorithm. The KNN with BPNN requires much less time for training and testing time as compared to all algorithms. If accuracy is considered then 98.92 is the accuracy percentage of KNN with BPNN algorithm which is more as compared to all algorithms. So because of this KNN with BPNN provides accurate diagnosis of liver patient data in quick time with respect to benign and malignant factor. A medical expert will find no difficulty while analyzing of this data to accurate treatment for liver disease patient. Following graph is there which shows the comparison of this algorithm with respect to accuracy and training and testing time parameter.



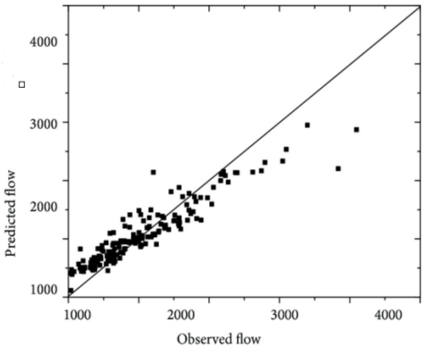
Graph 1. Comparison of techniques

As shown in above bar chart the training and training time gets on reducing from NBV to proposed technique and the accuracy on increasing from NBV to KNN with BPNN algorithm. From above chart it is clear that the above KNN with BPNN having much more accuracy than all state algorithm in given charts.

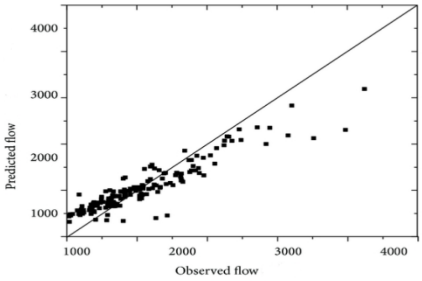


Graph 2. Line Chart to show comparison

Above line chart is used to show how training and testing time decreasing as the chart moves fromNBV to proposed algorithm and the accuracy goes on increasing from NBV to KNN with BPNN algorithm. In this way proposed algorithm KNN with BPNN is used for the diagnosis of data in quick time with maximum efficiency as compared to other algorithm.



Graph 3. Scatter plot of Comparison



Graph 4. Plot of Comparison

VI. Conclusion

In this technique combined algorithm of KNN and back propagation is used for the diagnosis of liver. It tests the percentage of benign and malignant in liver. Result is compared with previously used technique. The testing time and training time is improved by using this combined approach. The result from KNN and Back propagation are not so accurate. By combining them the result is obtained in quick time with maximum accuracy. The classification is done with KNN algorithm is back traced by using back propagation algorithm so that problem is resolved effectively by using this algorithm and due to that efficiency of this algorithm is improved. Error signal generated by back propagation algorithm helps to solve the deficiency of KNN algorithm. The conclusion of this technique is that diagnosis of disease is performed effectively with combined algorithm of KNN and Back Propagation algorithm.

VII. REFERENCES

[1]. R. A. Jacobs, "Increased rates of convergence through learning adaptation, Neural Networks", Elsevier, Vol.1 pp.295- 307, 1988.

[2]. Hoon Jin Seoungcheon and jinhong Kim, "Decision Factors on Effective Liver Patient Data Prediction",IJBB, Vol 6, 2014.

[3]. Lei Wang,Latifur Khan, and BhavaniThuraisingham, "An Effective Evidence Theory based K-nearest Neighbor (KNN) classification",WIIAT,IEEE,2008.

[4]. Saduf, MohdArifWani, "Comparative Study of Back Propagation Learning Algorithms for Neural Networks",IJARCSSE, 2013.

[5]. http://info.cancerresearchuk.org/cancerstats/types /liver/uk-liver-cancer-statistics.

[6]. http://www.cancer.net/patient/Cancer+Types/Liv er+Cancer?sectionTitle = Statistics.

[7]. http://www.cancerfoundationofindia.org/.

[8]. Paul R. Harper, "A review and comparison of classification algorithms for decision making", Elsevier,March 2005, Volume 71, Issue 3, Pages 315–331.

[9]. BendiVenkataRamana, Prof. M.Surendra Prasad Babu, Prof. N. B. Venkateswarlu," A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis", IJDMS, Vol. 3, 2011.

[10]. Daniel Hartono Sutanto and Mohd. KhanapiAbd. Ghani, "Improving Classification Performance Of K-Nearest Neighbour By Hybrid Clustering And Feature Selection For Non-Communicable Disease Prediction", ARPN Journal of Engineering and Applied Sciences, Vol. 10,Sept. 2015

[11]. Dr. K.Vijayarekha, "Back Propagation Neural Network", SASTRA University, Thanjavur,Lecture No-4, MHRD.

[12]. Michael J. Sorich, John O. Miners, Ross A. McKinnon,David A. Winkler, Frank R. Burden, and Paul A. Smith, "Comparison of linear and nonlinear classification algorithms for the prediction of drug and chemical metabolism by human UDP- Glucuronosyl transferase Isoforms" ACM, 2003, 43 (6), pp 2019–2024.

[13]. Prof.M.S.PrasadBabu, BendiVenkataRamana, Boddu Raja Sarath Kumar, "New Automatic Diagnosis of Liver Status Using Bayesian Classification",IEEE,Vol-2., PP-385-388,2010.

[14]. MirzaCilimkovic, "Neural Networks and Back Propagation Algorithm", 2015.

[15]. Shaveta Sharma1, Parminder Singh, "Speech Emotion Recognition using GFCC and BPNN", IJETT, Vol-18, PP-321-322, 2014.

[16]. S,Karthik, A.Priyadarshini, J.Anuradha, K.Tripathy," Classification and rule extraction using Rough set for the diagnosis of Liver Disease and its types", Advances in Applied Science Research,2011,2(3),334-345

[17]. Y.Unal, H.E.Kocer,H.E.Akkurt," Automatic Diagnosis of Intervertebral Degenerative Disc Disease using Artificial Neural Network",IATS"11, 16-18,May 2011

[18]. Fuzzy Logic- A practical Approach, F.MartinMc. Neil, Ellen Thro, Academic Press,1994

[19]. Fuzzy Expert System and Fuzzy Reasoning ,William Siler, James J Buckley,2005

[20]. Prof.Hua Li, Prof.Madan Gupta, Fuzzy Logic and Intelligent System, Kluwer Academic Publisher(1995).