AN AUTOMATED LIVER DISEASE DETECTION USING KNN

[1]Geeta Kadu, [2]Dr. Ranjana Raut, [3]Mr. Suraj S. Gawande
[1]Dept of Electronics Engineering, [2]Dept of Applied Electronics, [3]Application Engineer
[1]Dept of Electronics Engineering
[1]Shri Datta Meghe Polytechnic, Nagpur, Maharashtra, India

Abstract: Liver diseases are usually detected in the advanced stage. Hence, mortality rate is high in case of liver patients. Early diagnosis of liver diseases can definitely improve this. Diagnosis of normal and diseased liver is done using clinical data which is derived through blood tests. In this paper an attempt is made to make demarcation between the normal and diseased liver. Liver dataset by BUPA Medical Research Ltd., available on University of California at Irvine (UCI) repository, is used. Classification algorithm K-nearest-neighbor (KNN) is implemented for detecting the liver function status. This supervised classification techniques will train the algorithm to learn to differentiate between the normal and abnormal liver. Performance of KNN is evaluated using performance parameters accuracy, mean square error (MSE), sensitivity and specificity for 1 to 10 kernel functions. Performance for different kernel functions is compared. Result of this work is compared with the similar work done in previous research papers.

Index Terms - **K-nearest-neighbor (KNN), University of California at Irvine (UCI).**

I. INTRODUCTION

A healthy liver, shiny pinkish-brown triangle tucked under the right rib cage, is what we expect for every human being. Liver plays a vital role of removing toxins, converting digested food to energy, storing vitamins and minerals and controlling how much fat and sugar is sent back out to the rest of the body. Malfunctioning of liver can disturb functioning of other body parts. Hence, it is of utmost importance to detect the status of liver functioning at an early stage. Further investigations can be carried out once demarcation is done between normal and affected liver. Liver functioning can be diagnosed after checking the levels of various parameters given in the Liver Function Tests (LFT) in the pathological report of a patient. An automatic decision support system (DSS) can be of a great help in such cases for the correct diagnosis of a patient.

Doctors can rely upon the automatic classification methods as a diagnostic tool. In these tools information is derived by setting a correlation between large data sets. Most of the automatic classification methods work on the large data base of normal and affected patients [1]. During literature survey it was observed that various data mining techniques are used by the earlier researchers for the early detection of liver diseases from the pathological reports. Different data mining techniques such as Decision Tree, Support Vector Machine, Naïve Bayes, Artificial Neural Network (ANN) etc can be used for Diagnosis of Liver Disease [2]. Trained neural network with adaptive activation function. Opt-aiNET, an Artificial Immune Algorithm (AIS), was used for forming set of rules for liver disorder classification [3]. Decision Tree (DT), k-Nearest Neighbor (k-NN), Multi-Layer Perceptron (MLP), Naïve Bayes (NB), Logistic Regression (Logistic) and Random Forest Classification algorithms were implemented for liver disease diagnosis. Effectiveness of data mining algorithm is based on the characteristics of data [4]. Linear (LDA and DLDA), nonlinear (QDA, DQDA, NB and ANN) and decision tree (CART) classification algorithms were implemented for liver disorder diagnosis [5]. Computational intelligence techniques viz. J-48, MLP (Multi Layer Perceptron), Random Forest (RF), MLR (Multiple linear Regressions), *Support Vector Machine (SVM)*, Genetic programming (GP) techniques implemented on ILPD dataset for Liver Patient classification [7]. ANN based classifiers such as BP, RBF, SOM, SVM were implemented on selected attributes of ILPD data. Performance of all classifiers compared [8].

In this work, KNN is used as a classification algorithm for the demarcation of normal and diseased liver. Liver dataset by BUPA Medical Research Ltd., available on UCI repository, is used. KNN is used by implementing ten different kernel functions. Evaluation is done using performance parameters accuracy, mean square error (MSE), sensitivity and specificity. Best kernel function is selected after comparing performance parameters.

II. DATA DESCRIPTION

The work done in this paper used liver dataset by BUPA Medical Research Ltd. Available on UCI repository. The dataset includes 345 instances with 7 attributes. The first 5 attributes corresponds to blood tests which are thought to be very sensitive to liver disorders that might arise from excessive alcohol consumption. First five attributes are: mean corpuscular volume (mcv), alkaline phosphotase (alkphos), alamine aminotransferase (sgpt), aspartate aminotransferase (sgot), gamma-glutamyltranspeptidase (gammagt). The sixth attribute corresponds to number of half-pint equivalents of alcoholic beverages drunk per day (drinks). The last attribute represents selector field which indicates the state of the liver i.e. affected or not affected.

Out of 345 samples 200 are classified as value 1 and the remaining 145 are classified with value 2. Attributes and their description are given in table 2.1.

Table 2.1: Attributes and their description

Sr no.	Attribute	Description
1	mcv	mean corpuscular volume
2	alkphos	alkaline phosphotase
3	sgpt	alamine aminotransferase
4	sgot	aspartate aminotransferase
5	gammagt	gamma-glutamyl transpeptidase
6	drinks	number of half-pint equivalents of alcoholic beverages drunk per day
7	selector	field used to split data into two sets

III. METHOD

The whole design & development carried out in two stages: 1) Data processing 2) Implementation of classifier. Liver data from BUPA is processed for checking missing values. Data is separated between input and output. Columns 1 to 6 are selected as inputs and last column as output. 50 % data is given to the classifier. KNN is used as a classifier. Block diagram of the proposed system is shown in fig. 3.1 below:

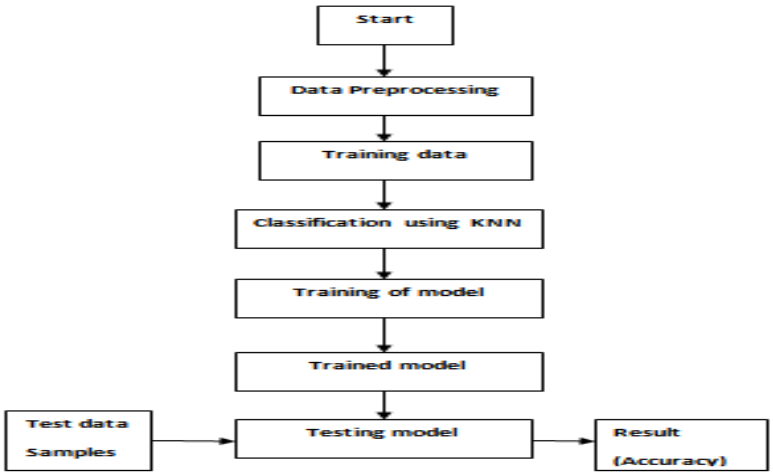

Figure 3.1. Block diagram of the system

3.1 KNN

K-Nearest neighbor algorithm (KNN) classifier is a non-parametric method used for both classification and regression. It is commonly used for classification in medical diagnosis. It is known for its simplicity and of interpretation and highly competitive results with low calculation time.

KNN is basically used for lower dimensional data i.e. less number of input variables and for a classification study when there is little or no prior knowledge about the distribution of the data.

KNN, unlike most of other supervised learning algorithms, does not learn any model. KNN makes predictions using the training dataset directly. It assumes that the training data is in a feature space. Testing data is also placed in this feature space. K decides the number of training instances that will remain in the periphery of the sample. To determine which of the K instances in the training dataset are most similar to a test sample a distance measure is used. The most popular distance measure is Euclidean distance.

The k-nearest-neighbor classifier calculates the Euclidean distance between the training samples and a test sample as given below:

$$d(\mathbf{x}_i, \mathbf{x}_l) = \sqrt{(x_{i1} - x_{l1})^2 + (x_{i2} - x_{l2})^2 + \cdots + (x_{ip} - x_{lp})^2}. \quad (3.1)$$

Where:
d(**x***i* , **x***l*) - Euclidean distance between sample **x***i* and **x***l*
x*i* -Input sample with *p* features (*xi*1,*xi*2,…,*xip*)
n - Total number of input samples (*i*=1,2,…,*n*)
p - Total number of features (*j*=1,2,…,*p*)

Hamming Distance, Manhattan Distance and Minkowski Distance are also other commonly used distance measures.

The value of K depends on the data. Increase in value of K reduces the effect of noise on the classification, but it also makes boundaries between the classes less distinct. Value of K can be selected between 1 to 21 and which K gives the best performance is checked.

3.2 EVALUATION

Performance parameters used for the evaluation of the selected machine learning classifier – KNN, are defined as follows:

Error rate: The error rate of a classifier is the percentage of the test set that are incorrectly classified by the classifier.

Error rate = (Incorrectly Classified Samples/ Classified Samples) × 100

Accuracy: Accuracy is the percent of correct classifications.

Accuracy = 1 - Error rate

OR

Accuracy = (TP+TN) / (TP+TN+FP+FN) (3.2)

Mean square error (MSE): MSE is the mean of the squares of the errors. Formula MSE is given below:

$$MSE = 1/n \sum_{i=1}^{n} \left(Yi - \hat{Y}_i \right)^2 \quad (3.3)$$

Sensitivity: Sensitivity is referred as True positive rate.

Sensitivity can also be defined as the ratio of Correctly Classified Positive samples and True Positive Samples as given below:

Sensitivity = TP / (TP + FN) (3.5)

Specificity is the true negative rate that is the proportion of true negative samples

Specificity can also be defined as the ratio of Correctly classified negative samples and True negative Samples as given below:

Specificity = TN / (TN + FP) (3.6)

Where:

TP - True Positive

TN - True Negative

FP – False positive

FN – False negative

$\hat{Y}$ - Vector of n predictions

Y –Vector of observed values of the variable being predicted

IV. RESULT & DISCUSSION

The developed application software efficiently performed for the classification of normal and abnormal liver. In this application software KNN is used. Performance of KNN is evaluated using performance parameters accuracy, MSE, sensitivity and specificity for the value of K from 1 to 10. Values of performance parameters are given in table 4.1 for K = 1 to 10.

For KNN, the best results are obtained for K = 1 i.e. Accuracy = 90.7%, MSE = 0.093 Specificity = 0.902 and Sensitivity = 0.9143.

Table 4.1. Performance of KNN for K = 1 to 10

Value of K	Accuracy	MSE	Specificity	Sensitivity
1	90.7	0.093	0.902	0.9143
2	89.7	0.095	0.892	0.902
3	74.42	0.2558	0.7647	0.71245
4	86.05	0.1395	0.8627	0.8571
5	68.02	0.3198	0.7255	0.61
6	81.98	0.1802	0.8431	0.7571
7	72.67	0.2733	0.7941	0.6286
8	78.49	0.2151	0.8627	0.6714
9	69.19	0.3081	0.7549	0.5857
10	76.16	0.2384	0.8137	0.6857

V. CONCLUSION

An automated Liver disease detection system is implemented using KNN. BUPA clinical data provided by BUPA Medical Research Ltd., available on University of California at Irvine (UCI) repository, is used. for classification. Output is in the form of two classes, hence KNN is used as supervised algorithm to learn to differentiate between the normal and abnormal liver. Performance of KNN is evaluated for 1 to 10 kernel functions. Highest performance was achieved for K = 1 i.e. Accuracy = 90.7%, MSE = 0.093 Specificity = 0.902 and Sensitivity = 0.9143.

Since numbers of input variables are less i.e.6 KNN is more convenient and easy to implement.

REFERENCES

[1] Mitra M., Mahdieh M., Amin B and Mohammad J. 2014. Identifying efficient clinical parameters in diagnose of liver disease. HealthMED - Volume 8 / Number 10.

[2] Harsha P. and Deepak Kumar X. 2016. A Survey on Diagnosis of Liver Disease Classification. International Journal of Engineering and Techniques - Volume 2 Issue 3, May – June.

[3] Humar K.and Novruz A. 2009. Mining Classification Rules for Liver Disorders. International Journal of Mathematics and Computers in Simulation. Issue 1, Volume 3.

[4] Hoon J., Seoungcheon K., and Jinhong K. 2014. Decision Factors on Effective Liver Patient Data Prediction. International Journal of Bio-Science and Bio-Technology. Vol.6, No.4, pp.167-178.

[5] Aman S. and Babita P. 2016. Liver disorder diagnosis using linear, nonlinear and decision tree classification Algorithms. International Journal of Engineering and Technology (IJET). Vol 8 No 5, Oct-Nov.

[6] Brett A. Lidbury, Alice M. Richardson and Tony Badrick. 2015. Assessment of machine-learning techniques on large pathology data sets to address assay redundancy in routine liver function test profiles. Published by publiDe Gruyter, Diagnosis. 2(1): 41–51.

[7] Jankisharan P., Rajan V., Jagdish M., and Sanjay P. 2014. Liver Patient Classification using Intelligence Techniques. International Journal of Advanced Research in Computer Science and Software Engineering. Volume 4, Issue 2, February.

[8] Anil Kumar Tiwari, Lokesh Kumar Sharma and G. Rama Krishna. 2013. Comparative Study of Artificial Neural Network based Classification for Liver Patient. Journal of Information Engineering and Applications. Vol.3, No.4.