Integrated Intelligent Research (IIR)
International Journal of Computing Algorithm
Volume: 06 Issue: 01 June 2017 Page No.37-40
ISSN: 2278-2397

A Tentative analysis of Liver Disorder using Data Mining Algorithms J48, Decision Table and Naive Bayes

P. Kuppan[1], N.Manoharan[2]
[1]Assistant Professor, Dept of Computer Science, Thiruvalluvar University College of Arts and Science, Thennangur, Vandavasi.
[2]Head and Assistant Professor, Dept of Computer Science, Thiruvalluvar University College of Arts and Science, Thennangur, Vandavasi
email: kuppan.anj@gmail.com, mano_n@rediffmail.com

Abstract — Nowadays healthcare field has additional data mining process became a crucial role to use for disease prediction. Data mining is that the process of investigate up info from the huge information sets. The medical information is extremely voluminous. Therefore the investigator is extremely difficult to predict the disease is challenging. To overcome this issue the researchers use data mining processing technique like classification, clustering, association rules so on. The most objective of this analysis work is to predict disease supported common attributes intake of alcohol, smoking, obesity, diabetes, consumption of contaminated food, case history of liver disease using classification algorithm. The algorithms employed in this analysis work are J48, Naive Bayes. These classification algorithms are compared base on the performance factors accuracy and execution time. The investigational results could be a improved classifier for predict the liver disease.

Keywords:— Datamining; classification; guided classification; Liver Disorder(j48,Naïve Bayes)

I. INTRODUCTION

The liver is that the largest organ of our body. It performs over than 500 completely different operate together with digestion, metabolism, storing nutrients and removing toxins and waste product. There are many disorders of the liver that need clinical care by a way of a physician and other healthcare professional as the records of patients clinical trial and other attributes are available electronically. The purpose of finding patterns in medical database is to identify those patients which share common attributes intake of alcohol, smoking, obesity, diabetes, consumption of contaminated food, family history of liver disease. In data mining, classification techniques are much popular in medical diagnosis and predicting diseases [1].

In this research work, Naïve Bays and support vector Machine (SVM) classifier algorithms are employed for liver disease prediction. There are several number of liver disorders that required clinical care of the physician [3]. The main objective of this research work is to predicate liver diseases such as cirrhosis, Bile Duct, Chronic Hepatitis, Liver cancer and Acute Hepatitis from Liver function test (LFT) dataset using above classification algorithms. The liver is the next largest internal organ in the human body, playing a major role in metabolism and serving several vital functions, e.g. Decomposition of red blood cells,etc.,.[7]

II. LITERATURE REVIEW

Dhamodharan etr.al [3] has predicted three major liver diseases such as liver cancer, Cirrhosis and Hepatitis with the help of distinct symptoms. They used naïve Bayes and FT tree algorithms for disease prediction. Comparison of these two algorithms has been done based on their classification accuracy measure. From the experimental results they concluded the Naïve bayes as the better algorithm which predicted diseases with maximum classification accuracy than the other algorithm. *Rosalina et al [13]* predicted a hepatitis prognosis disease using support vector machine (SVM) and wrapper method. Before classification process they used wrapper method to remove the noise features. Firstly SVM carried out feature selection to get better accuracy. Features selection were implemented to minimize noisy or irrelevance data. From the experimental results they observed the increased accuracy rate in the clinical lab test cost with minimum execution time. They have achieved the target by combining wrappers method and SVM techniques.

Omar S.Solaiman et al [10] has proposed a hybrid classification system for HCV diagnosis, using modified particle swarm optimization algorithm and least squares support vector machine (LS-SVM).Feature vectors are extracted using principal component analysis algorithm. As LS-SVM algorithm is sensitive to the changes of values of its parameters, Modified –PSO algorithm was used to research for the optimal values of LS-SVM parameters in less number of iterations. The proposed system was implemented and evaluated on the benchmark HCV data set from UCI repository of machine learning database. It was compared with another classification system, which utilized PCA and LS-SVM. From the experimental results the proposed system obtained maximum classification accuracy than the other systems. *Karthicket.al [*7] were applied a soft computing technique for intelligent diagnosis of liver disease .they have implemented classification and its type detection in three phases. In the first phase, they classified liver disease using Artificial Neural network (ANN)classification algorithm. in the second phase ,they generated the classification rules with rough set rule induction using learn by example (LEM) algorithm. In the third phase fuzzy rules were applied to identify the types of the liver disease.*ChaitraliS.Dangareet.al [2]* has analyzed prediction system for heart disease using more number of input attributes. The data mining classification techniques, namely decision tree, Naïve bayes and Neural Networks are analyzed on Heart disease database. The performances of these techniques are compared, based on accuracy. Another analysis shows that out

International Journal of Computing Algorithm
Volume: 06 Issue: 01 June 2017 Page No.37-40
ISSN: 2278-2397

of these three classification models neural networks has predicated the heart disease with highest accuracy.

III. ANALYSIS OF LIVER DISORDER

A. Risk Factors

Although alcohol is that the most typical reason cause of liver disease, it is not the only issue which will damage your liver. Here are measure another factors which will increase your risk of liver damage.

1)Diabetes: Having diabetes increases the possibility of liver disease by 50 percent. People who have diabetes as a result of insulin resistance have high quantities of insulin inside their blood that makes the abdominal weight gain. It causes the liver to store fat internally, causing fatty liver disease. Here all articles are measure customary connected to diabetes.

2).salt intake: High salt intake is acknowledge to cause cardiovascular disease, however it may also cause fatty liver disease by building up fluid within the liver (water retention) and swelling it up. Here are some helpful tips to scale back salt intake.

*3).Smoking***:** Although cigarette smoke does not want a primary impact on liver perform, harmful chemicals in cigarette smoke increase aerobic stress of the system once attaining the liver, inflicting irreversible damaging to the liver cells.

*4). Use of nutritional supplements***:** Dietary or biological process supplements will increase the assembly of sure liver enzymes once taken in excess amounts, inflicting damage.

5). Pesticides and heavy metals: Expertise of chemicals in pesticides and significant metals through vegetables, fruits and adulterate foods and can also harm the liver. These toxins get hold on in liver over a very long time to cause liver damage.

IV. EXTRACTION OF LIVER DISEASE DATA WAREHOUSE

The liver disorder data warehouse provides the screening the information of liver disorder patients. Initially the information warehouse is pre-processed to help make the mining process more efficient. In the paper WEKA tool can be used to compare the performance accuracy of data mining algorithms for diagnosis liver disease dataset .the pre-processed data warehouse is then classified using WEKA tool The feature selection in the liver disease. Using supervised machine learning algorithm such as for example Naive Bayes, Decision Table and J48 the end result are compared WEKA is an accumulation of machine learning algorithm for data mining tasks. The algorithm to be applied straight to a dataset. WEKA contains tools for data classification, Associate, Clustering and Visualization. It can also be suitable for developing new machine learning schemes. This paper concentrates on functional algorithms like naïve Bayes, Decision Table and J48.

1. Classification:

The fundamental classification is dependent on supervised algorithms. Algorithms are requested the input data .Classification is performed to understand the just how data will be classified. The classify tab can also be supported which shows the set of machine learning algorithms. These algorithms are generally operates on a classification algorithm and run it multiple times manipulating algorithm parameters or input data weight to improve the accuracy of the classifier .two learning performance evaluators are includes with WEKA. The every first simply splits a dataset into training and test data. While the next performs cross-validation using folds. Evaluation is normally described by the accuracy. The run information can also be displayed. For quick inspection of how well a classifier works.

2. Manifold machine learning algorithm:

The key motivation for different supervised machine learning algorithm is accuracy improvement. Different algorithm use different rule for generalizing different representation of the knowledge. Therefore, they have a tendency to error on different elements of the instance space. The combined utilization of different algorithms could result in the correction of the average person uncorrelated errors. As a result the error rate and time taken to produce the algorithms is compared with various algorithm

V. RESULT AND DISCUSSION

1. Classify

The consumer has the selection of applying sort of algorithms to the knowledge set which will the idea is that build a illustration of the knowledge used to produce observation easier. It is tough to identify that of the choices would supply the most effective output for the experiment. The most effective approach is usually to independently apply a combination of the on the market decisions and see what yields one thing around the specified results. The Classify tab is wherever a private selects the classifier decisions.

Again there area unit several choices to be elect at intervals the classify tab. Check choices provides the user the choice victimization four completely different check mode situation on the knowledge set: 1. Use training set 2. Supplied training set 3. Cross validation

2. Split percentage

There is the selection of applying any or each one of the modes to supply results which may be compared by the user. Additionally within the check choices toolbox there is a dropdown menu that the user will opt for varied things to use that with regard to the selection can give output choices like as an example saving the outcomes to file or specifying the random seed worth to be requested the classification.

3. Cluster

The Cluster tab opens the strategy that's wont to determine commonalties or clusters of occurrences at intervals info the knowledge set and turn out information for the user to research. There are always a couple of choices within the cluster window that area unit kind of like those with in the classifier tab. They are use training set, supplied test set, percentage split. The fourth choice is categories to cluster analysis that compares however well the information compares with a pre-assigned category at intervals the information. This can be often helpful if you can notice specific attributes inflicting the outcomes to be out of vary and for big data sets.

4. Associate

The associate tab opens a screen to settle on the alternatives for associations at intervals the information set. The user selects

International Journal of Computing Algorithm
Volume: 06 Issue: 01 June 2017 Page No.37-40
ISSN: 2278-2397

one in every of several decisions and method begins to yield the results.

5. Select Attributes

The following tab may be wont to opt for the actual attributes helpful for the calculation method. By default each one of the available attributes are utilized within the analysis of the information set. If the utilization desired to exclude sure style of the information they might deselect the particular decision from the list within the cluster window. It is helpful if a number of the attributes area unit completely different type like for instance alphanumeric data that may alter the results. The application searches through the chosen attributes to settle on that of them can best work the required calculation. To execute this, the user has to choose two options, an attribute evaluator and a research method. Once this is performed the program evaluates the information on the basis of the sub group of the attributes then performs the necessary seek out commonality with the data.

6. Visualization

The last tab with in the window is that the visualization tab. At intervals the program calculations and comparisons have occurred on the information set. Picks of attributes and ways in which of manipulation have already been chosen. The ultimate piece of the puzzle is gazing the information that has been derived throughout the method. The consumer are currently able to truly begin to ascertain the fruit of those efforts in a two dimensional illustration of the information.

The primary screen that the user sees once they choose the visualization choice is basically a matrix of plots representing the assorted attributes within the data set plotted against the opposite attributes. If necessary there is a scroll bar to look at all of the created plots. The consumer will opt for a precise plot from the matrix to see its contents for analization. A grid pattern of the plots permits the user to pick out the attribute positioning to their feeling and for higher understanding. Once a specific plot has been selected the consumer can remodel the attributes from one read to a different providing flexibility.

VI. EXPERIMENTAL SETUP

Naïve Bayes

A Naïve bayes classifier is just a simple probabilistic classifier based on applying bayes thermo with strong independent assumption. An even more descriptive term for the underlying probability model will be the self-determining feature model. In basic terms, a naïve bayes classifier assumes that the clear presence of a specific feature of a type of unrelated to the clear presence of some other feature [11]. The naïve bayes classifier performs reasonably well even though the underlying assumption is not true.

The advantage of the naïve bayes classifier used only requires a small amount of training data to estimate the means and variances of the variable necessary for classification because of independent variables are unspecified, only the variance of the variable for each label; need to be determined and not entire covariance matrix. In contrast to the naïve bayes operator, the naïve bayes (Kernel) operator can be applied on numerical attributes. This can be able in a clear –cut fashion using kernel density estimation and bayes' theorem:

$$V_{nb} = \text{argmax}_{v_j \in V} P(v_j) \prod P(a_i|v_j)$$

We generally estimate $P(a_i|v_j)$ using m-estimates:

$$P(a_i|v_j) = \frac{n_c + mp}{n + m}$$

where:

n = the number of training examples for which $v = v_j$
n_c = number of examples for which $v = v_j$ and $a = a_i$
p = a priori estimate for $P(a_i|v_j)$
m = the equivalent sample size

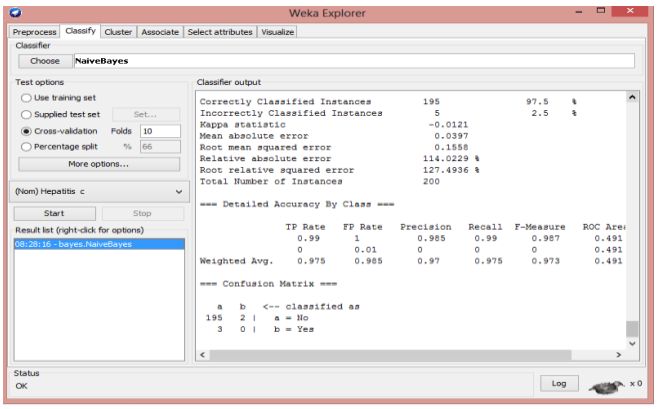

Fig1.This Screen shows the Naive bayes Algorithm result

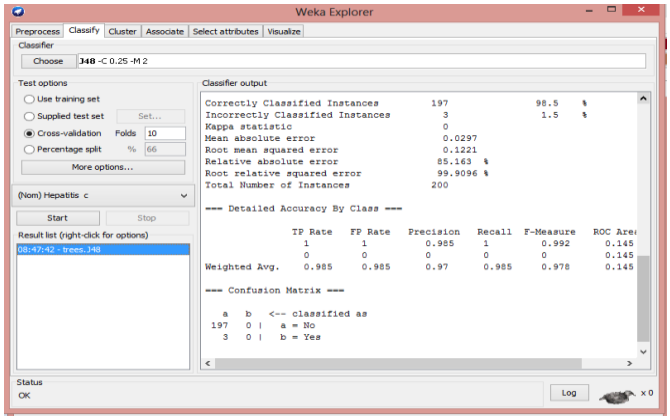

Fig2.This Screen shows the Output for J48 Algorithm result

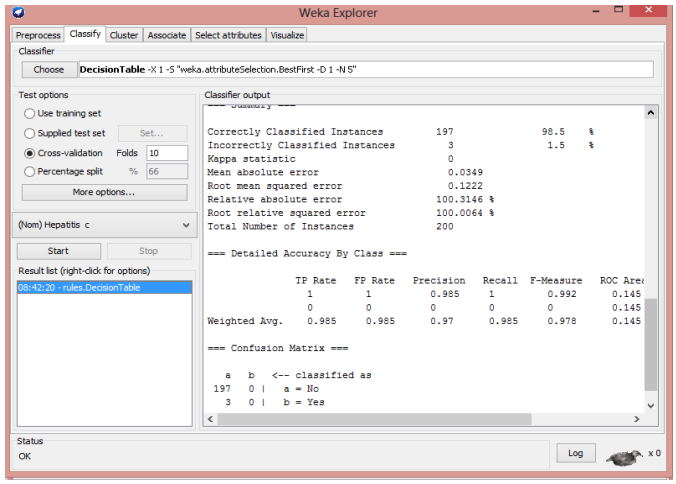

Fig3.This Screen shows the Output for Decision Table Algorithm

International Journal of Computing Algorithm
Volume: 06 Issue: 01 June 2017 Page No.37-40
ISSN: 2278-2397

Table1. Result for Naïve bayes, Decision table and J48 Algorithm

	NAIVE BAYES	DECISION	J48
ACCURACY	97.5%	98.5%	98.5%

Thus the research gives the most appropriate for predicting performance from the result. The J48 and Decision table perception gives 98.5% of accuracy prediction which is relatively higher than Naïve Bayes algorithm.

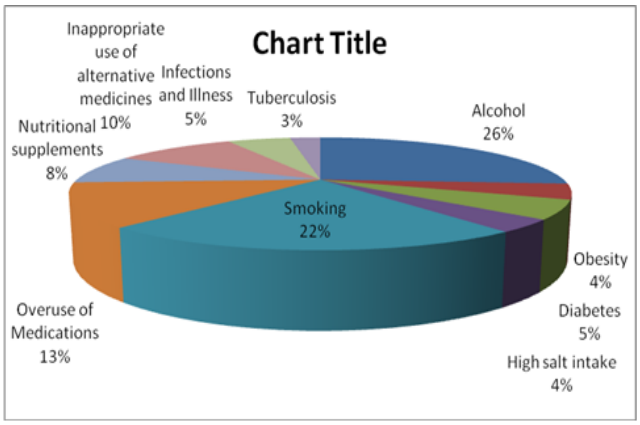

Fig4.This Figure shows the comparative analysis for the liver Disorder

VII. CONCLUSION

This research demonstrates the data mining based approaches that can be used to access factors of Liver Disorder Dataset. To presented an intelligent and effective Liver Disorder system using classification. Sample Data's are collected from some Hospitals who are all having Liver Disorder through for liver function Testing. This pre-processed data was classified with Classification algorithm using Weka tool. Using the Weka tool to perform the classification of data for the five main phases learning, experience and operation phases are carried out. Thus the system provides the way of comparative solutions given by various disorders we have got the result of liver Disorder which is having highest number of Liver Disorder with any sample collection. I proved that Liver Disorder is held by maximum number of people. This research the Liver disorder is mostly affected most of the people due to continuous smoking and highly intake of alcohol. The more number of male People is having more Liver Disorder than the female People. Overall survey shows between 35-65 Age People is highly affected in Liver Disorder. The Liver Disorder maximum Affected for Alcohol Habit People for. 26% of people are having their Liver Disorder, Smoking Habit People for. 22% of people are having their Liver Disorder,Obesity Patient People for. 4% of people are having their Liver Disorder, Diabetes Patient People for. 5% of people are having their Liver Disorder

References

[1] Bendi venkata Ramana, Surendra. Prasad Babu. M, Venkateswarlu. N.B,A critical study of selected classification Algorithms for Liver Disease Diagnosis, International Journal of Database Management Systems(IJDMS), Vol.3,No.2,May 2011 page no 101-114.

[2] Chaitrali S. Dangare, Sulabha S.Apte,"Improvved Study of Heart Disease prediction system using Data Mining Classification Techniques", International Journal of computer Applications (0975-888), Volume 47-No.10, june2012, pag3e no 44-48.

[3] Dhamodharan. S, Liver Disease prediction Using Bayesian classification, special Issue, 4th National conference on Advanced computing, Application, & Technologies, may 2014,page no1-3.

[4] Grimaldi. M, Cunningham. P,Kokaram. A, An evaluation of alternative feature selection strategies and ensemble techniques for classifying music, in: Work-shop in Multimedia discovery and Mining, ECML/PKDD03, Dubrovnik, Croatia, 2003.

[5] Gur Emre Guraksin, Huseyin Hakli, Harun U guz, Support vecto.r machines classification based on particle swarm optimization for bone age determination, Elsevier publications, Science direct, page no 597-602.

[6] Han,J.; Kamber,M., "Data Mining Concepts and Techniques". 2nd Edition, Morgan Kaufmann, san Francisco.

[7] Karthim. S, Priyadarishini. Anuradha. J and Tripathi. B.K, Classification and rule Extraction using Rough Set for Diagnosis of Liver Disease and its Types, Advances in Applied Science Research,2011, 2 (3):page no 334-345.

[8] Kotsiantis. S.B, Increasing the Classification Accuracy of Simple Bayesian classifier , AIMSA ,PP. 198-207,2004.

[9] Milan kumarai, Sunila Godara, comparative study of data mining Classification methods in cardiovascular Disease Prediction, International journal of Computer Science and technology, vol 2,Issue 2,June 2011, page no 304-308.

[10] Omar s.soliman,Eman abo Elhamd, Classification of Hepatities c virus using modified particle Swarm optimization and Least Squares Support Vector Machine, International Journal of Scientific & Engineering Research, volume 5,Issue 3,March -2014 122.

[11] Adachi U,Moore L.E, B Radford B.U,Gao W,Andthurman R.G Antibiotics to prevent liver injury in following long-term exposure to ethanol Gastroenterology 108:218-224,1995.

[12] Dufour M.C,S Tinson F.S,Andcaces M.F, Trends in cirrhosis morbidity and mortality seminars in Liver Disease 13(2):109-125,1993.

[13] Rosalina. A.H, Noraziah A. prediction of Hepatitis prognosis Using Support vector Machine and Wrapper Method, IEEE,(2010),2209-22.