International Journal of Engineering & Technology, 7 (3.34) (2018) 323-326

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper



Classification of Liver Patient Dataset Using Machine Learning Algorithms

S. Muthuselvan[1], S. Rajapraksh[2], K. Somasundaram[3], K. Karthik[4]

Aarupadai Veedu Institute of Technology, Vinayaka Mission's Research Foundation[1, 2, 3, 4]
Corresponding author E-mail: csmuthuselvan@gmail.com

Abstract

Prediction of the disease in the human being is the very long and difficult process in early days. Now a days, computer aided diagnosis is the important role in the medical industry for predicting, analyzing and storing medical information with the images. In this paper will discuss and classify the liver patients with the help of the liver patient dataset with the help of the machine learning algorithms. WEKA is the software used here for implement the some of the classification algorithms with the data selected from the liver disease dataset. After the successful implementation of the all the algorithms, the best algorithms selected from the output of the all the algorithms execution.

Keywords: *Liver, Classification, Naïve Bayes, R48, Random Tree, K-star.*

1. Introduction

The largest organ in an abdoman is the liver in the shape of triangular. The two parts of the livr is left and right hemi liver. It is a single organ. Liver used to essential for function our body. This is the primary organ for maintaining the chemicals like glucose, balancing the so many nutirents, fat, vitamis, choleserol and hormones.[1] In an early stage of the liver problem diagnosition will increase the the survival rate of the patient. Analyzing the enzymesin's levels will lead to diagnose the liver diseases from the blood[3].

In Knowledge Discovery Process, Data mining methods are partitioned into two noteworthy classes. These are expressive compose and forecast write. Every one of the sort will have distinctive kind of the methodologies[8]. Information Mining is a procedure of separating conceivably helpful, already obscure data from the crude information. Information mining is one stage in the KDD procedure. It is the most explored piece of the procedure. Information mining is characterized as a "sort of database investigation that endeavors to find valuable examples or connections in a gathering of information. The examination utilizes progressed measurable strategies, for example, bunch investigation, and in some cases utilizes counterfeit consciousness or neural system methods. A noteworthy objective of information mining is to find beforehand obscure connections among the information, particularly when the information originate from various databases [9]."Classification algorithm that is generally utilized as a part of expectations basing on chronicled information. Classification is a class expectation system, which is regulated in nature. This system has the capacity to anticipate the name for classes, gave that adequate quantities of preparing cases are accessible [10].

2. Background

2.1 Data Mining

Data mining centers on the disclosure of already obscure properties in the information. It needn't bother with a particular objective from the space, however rather centers on finding new and fascinating learning. Data mining is ideas are utilized to examination the information in complex way to find the data which not known as of now with the fascinating examples and to discover the connections among the tremendous volume of datasets [12].

2.2 WEKA

The goal of the WEKA software is to deliver an entire group of data pre-processing and machine learning algorithms for only for investigate researchers and the educationalist. This software will help us to think about the distinctive kinds of machine learning and data mining strategies. WEKA tool is actual easy to use, since, the tool was established as a modest Application Programming Interface using the Java language [13].

WEKA software fully based on the Graphical User Interface (GUI) chooser. This will be very useful to the user those who in the beginning stage of learning. Also it will reduce the time for using the software, because of its GUI concepts. WEKA contains the four applications combined together in single software. The application of the WEKA is Explorer, Experimenter, knowledge Flow and the Simple CLI.WEKA compatible with the medium size of the data. Application will be slow down, if the size of the data increased.

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

3. Classification Algorithms

3.1 Naive Bayes Algorithm

In this paper Naïve Bayes algorithm is the one of the algorithm used to implement the selected database. Naïve Bayes algorithm based on the Bayes theorem. The equation of this theorem is as follows,

$$p(c/d) = \frac{p(d/c)p(c)}{p(d)}$$

Where $p(c/d)$ is the probability of instance d being in class c, $p(d/c)$ is the probability of producing instance d given c, $p(c)$ is the probability of occurrence of class c, p (d) is the probability of instance d occurring. [4]. The advantage of this Naïve Bayes algorithms is, need small amount of the data needed for predicting the problems.

The confusion matrix of this algorithm can be constructed with the help of the incorrected classified as well as the corrected classified details. The confusion matrix is as follows,

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Where TP = total number of true positive in a samples taken, TN=total number of true negative in a samples taken, FP=total number of false positive in a samples taken and FN=total number of false negative in a samples taken from the dataset.

3.2 K-Star

The K* algortihm developed by John G. Cleary and Leonard E. Trigg in the year 1995. This algorithm will be the best one for finding the missing values, involving with mixed values and smoothness problems[5]. K-star algortihms is the instance based for the classifying the databases. The function of K-star is defined as follows,

$$K^*(b/a) = -\log_2 P^*(b/a)$$

The K* is not firmly as distance function.For example, $K^*(a/a)$ isin general non-zero and the function. The K* algorithm provable for the following properties,[5]

1. $K^*(b/a) \geq 0$.
2. $K^*(c/b) + K^*(b/a) \geq K^*(c/a)$

3.3 J48

The J48 decision tree implementing the Quinlan's C4.5 algorithm for making the C4.5 pruned or unpruned decision tree. The next level of Quinslan's ID3 algorithms is C4.5 [6]. The J48 algorithm used for the classifying the datasets for the process of data mining and the machine learning algorithms. J48 used to take a decision for the data splitting into the reduced subsets. The attributes which are discrete and the continuous, missing attributes and the training data can be handling by the J48. All these attributes are differing its values [7].

3.4 Random Tree

Random Tree is a managed Classifier; it is a group learning calculation that produces numerous individual learners. It utilizes a sacking thought to deliver an irregular arrangement of information for developing a choice tree. In standard tree every hub is part utilizing the best split among all factors. In a Random Forest, every node is part utilizing the best among the subset of predicators arbitrarily picked at that node [11].

4. Methodology

4.1 Dataset Selection

The data selection is the primary and important steps in the data mining and machine learning algorithm implementation. The selected tool for this dataset is good for the average value of the instance. If the number of instance in the dataset is very huge, the performance of these tools is not good. In this paper we have selection the Liver Disease dataset with the eleven parameters. [2] All these data are collected from the Andhra Pradesh's North East area in India. There are totally ten attributes are available in this dataset.

Table 1: List of Parameters with their data types

S. No.	Parameter Name	Data Type
1	age	Integer
2	gender	String
3	tot_bilirubin	Real
4	direct_bilirubin	Real
5	tot_proteins	Integer
6	albumin	Integer
7	ag_ratio	Integer
8	sgpt	Integer
9	Sgot	Real
10	Alkphos	Real
11	is_patient	Integer

The entire dataset having the 416 liver disease data as well as 142 non-diseased data available. The number of male patient is 441 and the 142 female patients are taken for our dataset. The table 2 describe about the list of attributes used in the application and the description of the same. The dataset belongs to north east, Andhra Pradesh in India. All these data are divided into two categories like liver patient and non-liver patient. Also these contain 441 male and 142 female patient data. If the patient age exceeds 90, here we have mentioned as 89.

Table 2: List of Attributes with the Description

S. No.	Attribute Name	Attribute Description
1	age	Age of the patient
2	gender	Gender of the patient
3	tot_bilirubin	Total Bilirubin
4	di-rect_bilirubin	Direct Bilirubin
5	tot_proteins	Total Proteins present in patient
6	albumin	Albumin amount of the patient
7	ag_ratio	Albumin and Globulin Ratio
8	sgpt	Alamine Aminotransferase
9	sgot	Aspartate Aminotransferase
10	Alkphos	Alkaline Phosphatase
11	is_patient	Whether the data is belongs to Liver disease patient or not

5. Architecture for Classification of Liver Patient Dataset

The architecture for Classification of the Liver patient dataset is shown in fig. 1. This flow is initiated from the dataset selection. In this paper .csv format dataset are selection for the implementing the algorithm. CSV format is nothing but comma-separated values. The alternative file format of this process is ARFF. The above mentioned two types of file format can be used for WEKA applications for executing the dataset. After completion of the data selection, all the data are completely pre-processed including the

missing values checking. Data discretization is performed for all the attributes for grouping the values based on the similarities and make them as single group. The next process is to implement the algorithm with the selected parameter. In this liver patient dataset, we have selected the parameter called "is_patient". This parameter values are one and two. The value one is represented as liver patient detail and two is represented as not a liver patient.

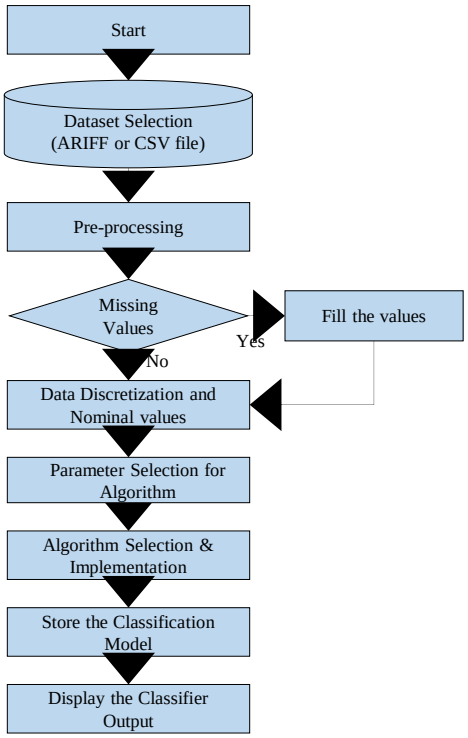


Fig. 1. Process flow diagram for Liver patient dataset classification

After the selection of the parameter, the algorithms are selected for executing the existing dataset. The result of this classification model was saved as separately. The output can be displayed separately as a classifier output.

6. Exploration &Pre-processing the Dataset

The data exploring is the primary stage of the evaluation of the algorithm in machine learning process. The dataset selection may be in any of the following file format. The data can be imported from the files C4.5, ARFF, CSV and binary. In this paper we are used the file format call CSV (Comma Separated Values).

Pre-processing the selected dataset is the important for the algorithm execution in WEKA application. The pre-processing tools are data discretization, data normalization, resampling, and attribute selections. Pre-processing is also called filter the dataset. In this paper, discretization was performed for the all attributes. The discretization will do the grouping for the distinct values and all the attribute values will convert into nominal values from the integer values. After completion of discretization, the "numeric to nominal values" filter also performed. The figure 2 shows the output of the database after completion of the few pre-processing steps with the help of the filter option available in the WEKA tool.

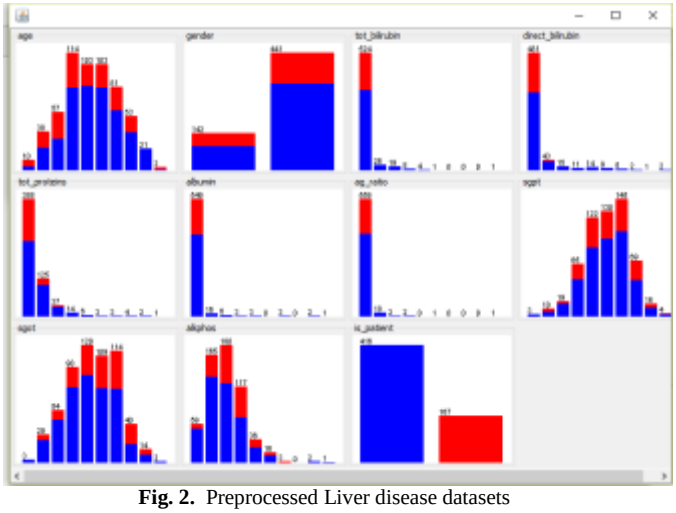


Fig. 2. Preprocessed Liver disease datasets

The above figure 2 shows the visualized report of after the completion of the discretization and the numeric into nominal values. This process will help to make the all distinct values into the single group as well as the conversion of the all values into nominal. The nominal values are very much compatible with the WEKA tool.

7. Algorithm Selection & Evaluation

The implementation stage of this proposed system is algorithm selection and evaluation. Data mining and machine learning algorithms are used to classification rule; clustering rule mining and Association rule mining are available in the WEKA application. In this liver disease data set, taken few classification algorithms for implementing the dataset. Naïve Bayesian algorithm, K-Star, Random Tree and J48 algorithms are selected for implementing the dataset. All these algorithms are executed with the help of the WEKA application. After completed the execution of the entire algorithm, the confusion matrix of these entire algorithm will be analysed and discussed.

The Naïve Bayes algorithm selected from the classifier tab in the WEKA explorer. The test options will be finalise before execution of the algorithm. Here, we have selected the percentage split as 66%. The classifier evaluation options also verified with the options for "output model" and "output confusion matrix". The performance study of all the algorithms are as follows in the table IV

8. Results and Discussion

The classification algorithms confusion matrix in our study is shown in the table 2. This is predicted into two categories like diseased dataset and the Non-disease dataset. All the data are combined together as a whole dataset. Using this dataset, applying the algorithm which is available in the WEKA classification is classified with the help of this confusion matrix.

Table 3: Confusion Matrix in Our Study.

Actual	Predicted	
	Disease (positive)	No-disease (negative)
Positive	TP	FP
Negative	FN	TN

The positive and negative predictions are categorized with the prediction is True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). The Table IV shows the performance study of the selected algorithm for the implementation of the selected datasets. All the four algorithms are executed for getting the accuracy and time taken for the execution time of the all algorithms.

Table 4: Performance Study of Algorithm

Algorithm Used	Accuracy (%)	Time Taken (Secs.)
Naïve Bayes	60.6%	0.02
K-star	67.2%	0
J48	71.2%	0.05
Random Tree	74.2%	0.05

The accuracy of the Naïve Bayes algorithm for the liver disease dataset is 60.6%, K-star is 67.2%, J48 is 71.2% and the Random Tree algorithm is 74.2%. This Random tree algorithm is highest accuracy with the all other algorithms with the highest time taken for the execution of the selected Datasets. The time takes for these all four algorithms are measure in the unit of seconds.

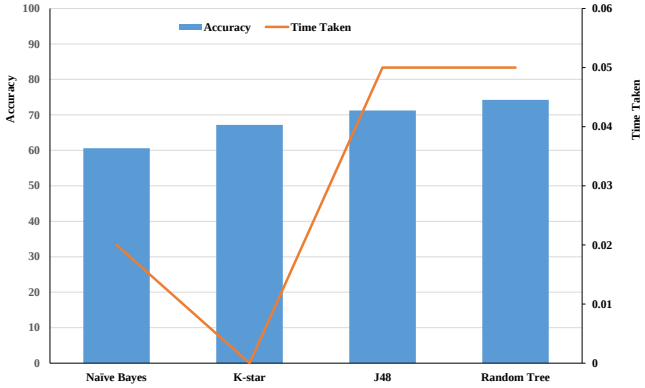


Fig. 3. Performance Study of Algorithm

The time taken for the execution of the all the algorithms are found from the executed datasets. The very least time taken for the algorithm is the K-star, which very smaller than the one second. The next highest time taken is Naïve Bayes 0.02secs which is higher compared to K-star. The J48 and the Random tree algorithms are equal with the 0.05 seconds. The above figure 3 shows the graphical representation of the performance study of the all four algorithms.

9. Conclusion

This persistence of this study is to assess and examined the data composed from the Andhra Pradesh's North East area in India for applying the machine learning algorithm with the help of the machine learning Tool. All the selected algorithms are executed and classified the liver disease patients as well the non-diseased patient from the selected data. The highest accuracy given with nominal execution time taken is the Random Tree algorithm. The time taken for the executing this algorithm is higher than the all other selected algorithms for this study. But the comparison between the all other algorithm, the Random Tree algorithm given higher accuracy. The liver disease predicted with the few machine learning algorithms. This study will helpful to the medical area people for the easy predictions.

References

[1] Parminder Kaur1and Aditya Khamparia, "Classification Of Liver Based Diseases Using Random Tree", *International Journal of Advances in Engineering & Technology*, June, 2015, ISSN: 22311963, Vol. 8, Issue 3, pp. 306-313

[2] Lichman, M. (2013). *UCI Machine Learning Repository* [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[3] Bendi Venkata Ramana, Prof. M.Surendra Prasad Babu,, Prof. N. B. Venkateswarlu,"A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis", *International Journal of Database Management Systems (IJDMS)*, Vol.3, No.2, May 2011.

[4] Chuan Choong Yang, Chit Siang Soh and Vooi Voon Yap, "A non-intrusive appliance load monitoring for efficient energyconsumption based on Naive Bayes classifier", *Sustainable Computing: Informatics and Systems* 14 (2017) 34–42.

[5] Cleary, J. and L. Trigg, "K*: An Instance-based Learner Using an Entropic Distance Measure", in *12th International Conference on Machine Learning*. 1995. p. 108-114.

[6] Ross J. Quinlan: "Learning with Continuous Classes" *In Proceedings AI'92 (Adams & Sterling, Eds)*, 343-348, Singapore: World Scientic, 1992.

[7] Youvrajsinh Chauhan and Jignesh Vania, "J48 Classifier Approach to Detect Characteristic of Bt Cotton base on Soil Micro Nutrient", *International Journal of Computer Trends and Technology (IJCTT)* – volume 5 number 6 –Nov 2013.

[8] S. Muthuselvan and Dr. K. Soma Sundaram, "An Analysis of Knowledge Discovery Process Over a Cloud Environment — A Survey", *International Journal of Applied Engineering Research* ISSN 0973-4562 Volume 10, Number 17 (2015).

[9] Inderjit Kaur, Deep Mann,"Data Mining in Cloud Computing", *International Journal of Advanced Research in Computer Science and Software Engineering*.

[10] Emmanuel Ahishakiye, Elisha Opiyo Omulo, Danison Taremwa and Ivan Niyonzima, "Crime Prediction Using Decision Tree (J48) Classification Algorithm" *International Journal of Computer and Information Technology* (ISSN: 2279 – 0764) Volume 06 – Issue 03, May 2017.

[11] Sushilkumar Kalmegh, "Analysis of WEKA Data Mining Algorithm REPTree, Simple Cart and RandomTree for Classification of Indian News", *International Journal of Innovative Science, Engineering & Technology* (ISSN: 2348 – 7968), Vol. 2 Issue 2, February 2015.

[12] S. Muthuselvan, Dr. K. Soma Sundaram and Dr. Prabasheela, *International Conference On Information Communication And Embedded System* (ICICES 2016), ISBN: 978-1- 5090-2552- 7.

[13] M. Hall, et al., "The WEKA data mining software: an update," *SIGKDD Explorer Newsletter*., vol. 11, pp. 10-18, 2009.