Case-Based Reasoning (CBR) with Hyper Lift for Classification of Hepatitis C Virus

[1]K Akshay Reddy,[2]G. Karthika
[1]UG Student, [2]Assistant Professor
[1, 2] Computer Science & Engineering, GITAM Institute of Technology
[1, 2] GITAM, Visakhapatnam, India

Abstract: Hepatitis, being a widespread and one of the most dangerous liver diseases, is affecting millions of people around the globe. Although, Hepatitis C virus (HCV) may be present in many other body parts, the target is liver. An effective diagnosis system based on Case-Based Reasoning (CBR) with hyper-lift metric for hepatitis C virus (HCV) is presented here. Data mining which is an efficient technique can play an important role in healthcare reform. Using the proposed algorithm, the presence of hepatitis C virus can be predicted in an efficient and precautionary way which helps immensely in early diagnosis of the infection.

Index Terms **– Hepatitis, Case-Based Reasoning, Data mining.**

I. Introduction

Hepatitis, one of the most dangerous diseases, is a viral infection that causes inflammation of the liver, that can be caused by various sources consisting of viral infections A, B, C, and also autoimmune hepatitis, faulty liver hepatitis, spirituous hepatitis, and toxin-induced hepatitis. World Health Organization (WHO) suggest that around 170 million people are affected by hepatitis C and 400 million people by chronic hepatitis B, of which 12.2 million HCV carriers are from India [1-3].

1.1 Hepatitis C

Hepatitis C virus was discovered in the year 1989 by Choo and co-workers and a widespread research and study on this viral epidemiology has begun ever since. Although, it may be present in many other body parts, its primary target is the liver. However, liver damage is caused by the reaction of the immune system to the virus and by HCV itself. The challenge with HCV infection is that acute hepatitis C shows negligible symptoms, which make early diagnosis highly difficult, and gradually develops into chronic hepatitis. After thorough research, the primary causes for HCV infections have been classified as blood contamination, sexual intercourse and re-usage of syringes [3]. The below table represents the percentages of HCV-infected patients across various countries. The graphical representation in pie chart i.e. Figure 1 for table is shown below.

Table 1 Percentage of HCV patients in various countries and rest of the world

Countries	Percentage (%)
Japan	1.2
India	1.5
Malaysia	2.3
Philippines	2.3
USA	2.5
Egypt	14.5
Rest of the world	75.5

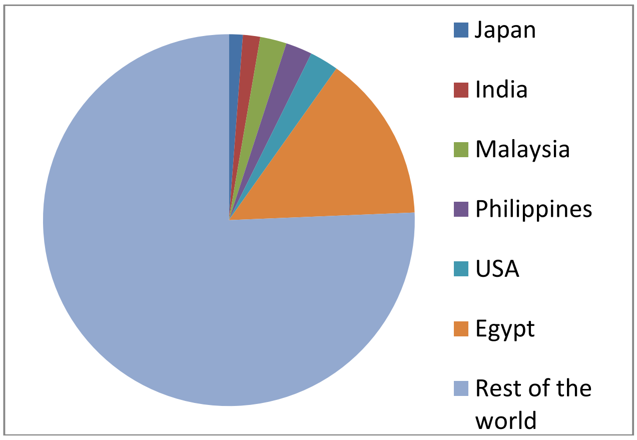

Figure 1 Pie chart representation of percentage of HCV-infected patients in various countries and rest of the world

1.2 Data Mining

In order to identify patients affected by HCV, and to enhance the accuracy of early diagnosis, data mining techniques can be used. Data mining is a method of understanding information, deducing useful hidden patterns and relationships in large data sets [4]. Using numerous algorithms predictive models can be developed with respect to the task [5].

In this paper, a new approach is proposed for the classification of patients and the most effective symptom, which helps in early diagnosis and treatment. The proposed algorithm is developed with the aid of Case-Based Reasoning (CBR), successfully, with the usage of correlation 'hyper lift' metric. First, the missing attributes in the data set are filled by applying 'statistical mode to the remaining values of the attribute after which hyper lift for every attribute with respect to the class attribute is calculated for each value. Now a new tuple is considered and compared with existing tuples to check for the identical training case. For non identical cases it searches for the most similar tuples. The proposed algorithm was implemented using R programming language on R Studio 1.0.136.

II. DATA DESCRIPTION

The hepatitis C data set , used to classify whether the patient will be alive or dead was taken, for the development of proposed algorithm , from UCI Machine Learning Repository, with the attribute values derived after carrying out several medical tests. It consists of 155 samples containing 19 attributes belonging to two different classes (live or die). There are 13 binary attributes and 6 numerical attributes in the data set, which are depicted below:

1. CLASS: DIE, LIVE
2. AGE: 10, 20, 30, 40, 50, 60, 70, 80
3. SEX: male, female
4. STEROID: no, yes
5. ANTIVIRALS: no, yes
6. FATIGUE: no, yes
7. MALAISE: no, yes
8. ANOREXIA: no, yes
9. LIVER BIG: no, yes
10. LIVER FIRM: no, yes
11. SPLEEN PALPABLE: no, yes
12. SPIDERS: no, yes
13. ASCITES: no, yes
14. VARICES: no, yes
15. BILIRUBIN: 0.39, 0.80, 1.20, 2.00, 3.00, 4.00
16. ALK PHOSPHATE: 33, 80, 120, 160, 200, 250
17. SGOT: 13, 100, 200, 300, 400, 500
18. ALBUMIN: 2.1, 3.0, 3.8, 4.5, 5.0, 6.0

19. PROTIME: 10, 20, 30, 40, 50, 60, 70, 80, and 90
20. HISTOLOGY: no, yes.

III. METHODOLOGY

This works uses correlation hyper lift metric for identifying the missing attributes in Case-Based Reasoning (CBR) with the statistical mode. Moreover, a new tuples is taken to check its redundancy and if exists then the result of the tuples is updated else hyper lift metric is used to predict.

3.1 Case-Based Reasoning

Case-Based Reasoning (CBR) uses complex symbolic illustrations to solve tuples for which they are stored and are highly functional for patient's case information and treatment to detect new patient [6].

3.2 Correlation

A correlation measure or metric is used to expand the support-confidence framework for association rules. There are several metrics in which one can be used and 'hyper-lift' metric as one among them. This is more robust for low counts using a hyper geometric count model [7].
The hyper lift metric is determined as follows:

$$hyperlift(X \Rightarrow Y) = \frac{C_{XY}}{Q_{\delta}[C_{XY}]} \qquad (1)$$

where X and Y are item sets.
C_{XY} is the number of transactions containing X and Y
$Q_{\delta}[C_{XY}]$ is the quantile of the hyper geometric distribution with parameters C_X and C_Y given by δ (typically the 99 or 95% quantile).
Range: [0,∞] (1 indicates independence)

The implication is that higher the value of the hyper lift, stronger the relation among the attributes. The proposed algorithm mainly deals with hepatitis C virus infection where identification of the attribute with the highest influence is calculated by hyper lift metric for every attribute with the class attribute to help in the discovery of which class (live or die) the patient belongs to in early diagnosis.

3.3 Algorithm

Step 1:
The raw data given in the text format is converted to comma separated values (CSV) format.
Step 2:
The respective modes and attributes of the dataset are used to fill the missing values.
Divide the given data set into two sets vis-à-vis training and test sets.
Step 3:
The probability of occurrence of each attribute value is calculated
Step 4:
Hyper Lift metric for each value of the attribute with respect to the class attribute values is calculated using the Eq. (1).
Hyper Lift can be greater than 1, less than 1, or equal to 1.
Step 5:
A new tuple from the test data is taken, and the attribute values of the new tuple are compared with the attribute values of all the tuples in the training set.
Step 6:
The hyper lift of a particular attribute value is considered for a tuple I in training set, if a match occurs in Step 5. The values of the lifts are added and stored in the variable 'lift_total' if the lift is greater than 1.
If a match does not occur and the lift value of that attribute value is greater than 1, the corresponding lift values are subtracted from the variable *'lift_total'*.
Step 7*:*
Repeat Step 6 for all the tuples in the training set.

The tuple in the training set with the maximum hyper lift value is considered as the one which is similar with the tuple from the test set.
Step 8:
The corresponding class attribute value of the similar tuple in the training set is given as the predicted class value for the tuple from the test set.
Step 9:
Repeat Steps 5, 6, 7, and 8 for all the tuples in the test set to predict their values.

IV. EXPERIMENTAL RESULTS

The accuracy of correctly predicting that a patient will die is greater than the accuracy of correctly predicting that the patient will live. The classifier conservatively predicts that the healthy patient may be prone to the HCV infection based on the reported symptoms which will help in early diagnosis and reduce the possible risks [6].
The measure of proportion of correctly identified positives, also called sensitivity, is called Recall

$$\text{Recall} = \frac{\#\,of\ True\ Positives}{\#\,of\ Positives} \qquad (2)$$

The correctly classified ‘live’ tuples are true positives.

The measure of proportion of correctly identified negatives is called Specificity.

$$\text{Specificity} = \frac{\#\,of\ True\ Negatives}{\#\,of\ Negatives} \qquad (3)$$

The correctly classified ‘die’ tuples are true negatives. The fraction of instances that are retrieved and relevant is called Precision.

$$\text{Precision} = \frac{\#\,of\ true\ positives}{\#\,of\ true\ positives + \#\,of\ false\ positives} \qquad (4)$$

The incorrectly classified ‘live’ tuples are false negatives.
The incorrectly classified ‘die’ tuples are false positives.

The percentage of tuples of the given test set that the classifier classified correctly is called Accuracy.

$$\text{Accuracy} = \text{recall}\left(\frac{pos}{pos+neg}\right) + specificity\left(\frac{neg}{pos+neg}\right) \qquad (5)$$

Here, pos denote the number of positive tuples and neg denotes the number of negative tuples.

The proposed algorithm returned an accuracy of 78% with performance measures: specificity, recall, and precision with values 0.88, 0.65, 0.92, and 0.77, respectively. The confusion matrix for the obtained results is given in Table 2, and the performance measures are shown in Table 3.

Table 2 Confusion matrix for hepatitis C virus data set

	Predicted			
Actual	**Classes**	Live	Die	% Correct
	Live	23	12	65.7
	Die	2	16	88.9

	Overall %			78

Table 3 Performance measures for the proposed algorithm

Accuracy measure	Value
Specify	0.88
Recall	0.65
Precision	0.92
Accuracy	78%

IV. CONCLUSION

Diagnosing disorders in one of the most difficult jobs for a physician as a mistake may lead the patient to death bed. In this regard, several data mining techniques are used to help the physician better diagnose the patient. In this paper, correlation metric—'hyper lift' in Case-Based Reasoning (CBR)—is used to diagnose hepatitis C virus (HCV)-infected patients. The missing attributes in the training set will be filled and the correlation metric 'hyper-lift' is calculated for every attribute value with the corresponding class attribute value. When a new patient's (tuple) attributes will be considered for classification using lift metric, it will check for a similar tuple in the training set. The accuracy of predicting which class a patient belongs to is 'die' class 88.9% and 'live' class 65.7 % which infers that a healthy person may be prone to HCV infection and should be alert regarding the infection. The overall accuracy by proposed algorithm is 78%. The proposed algorithm reduces the risk for the patients with the HCV symptoms by predicting which class they belong to. The proposed method will be instrumental in further research in the area of hepatitis disease diagnosis.

Despite the fact that a patient is not infected with HCV, the algorithm conservatively labels the patient as an HCV-infected patient considering danger to the patient based on their condition. This led to a higher number of false positives which reduced the accuracy. However, the patients who are truly infected with HCV are correctly labelled so with an accuracy of 89%. This will enable the patients who are not actually infected to take suitable precautionary measures.

V. ACKNOWLEDGMENT

We would like to convey our humble gratitude towards Mr. Vikas B, Assistant Professor, Computer Science and Engineering, for the encouragement and resources provided by him to complete this work.

REFERENCES

[1] WHO, Hepatitis C (Fact Sheet No. 164), World Health Organization, Geneva, 2000

[2] Hodgson S., Harrison R. F., Cross S. S., An Automated Pattern Recognition System For The Quantification Of Inflammatory Cells in Hepatitis-C-Infected Liver Biopsies, Image And Vision Computing 24, 2006, Pp. 1025–1038.

[3] Vikas B, Yaswanth D.V.S., Vinay W., Sridhar Reddy B., Saranyu A.V.H. Classification of Hepatitis C Virus Using Case-Based Reasoning (CBR) with Correlation Lift Metric. *4th International Conference on Information System Design And Intelligent Applications, 2017.*

[4] Huda Yasin, Tahseen A. Jilani, Madiha Danish.: Hepatitis-C Classification using Data Mining Techniques. International Journal of Computer Applications (0975 – 8887) Volume 24– No.3, 2011.

[5] AbuBakr Awad, Mahasen Mabrouk, Tahany Awad, Naglaa Zayed, Sherif Mousa, Mohamed Saeed.: Performance Evaluation of Decision Tree Classifiers for the Prediction of Response to treatment of Hepatitis C Patients. Pervasive Health, Oldenburg, Germany 2014.

[6] Kamber, Micheline, and Jian Pei. Data Mining, Second Edition. 1/e. Morgan Kaufmann, 2006. Print.

[7] Michael Hahsler and Kurt Hornik. New probabilistic interest measures for association rules. Intelligent Data Analysis, 11(5):437--455, 2007