A Comparative Study of Data Mining Techniques for HCV Patients' Data

[1]Tahseen A. Jilani, [2]Muhammad Shoaib, [3]Rehan Rasheed, [4]Bilal ur rehman

Department of Computer Science, University of Karachi, tahseenjilani@uok.edu.pk, , mshoaibuok@gmail.com, smrehanrasheed@gmail.com, bilal.ur.rahman@hotmail.com

Abstract

Hepatitis C is one of the most widespread sources of the liver failure and cancer and represents a major public health problem. Data mining techniques play's significant role in the field of Health informatics. Therefore we have applied different data mining techniques which include Naïve Bayesian Classification, Decision Tree and Fuzzy C-means on hepatitis C patients' data for observing the factors of high prevalence of the risk of hepatitis C virus. The dataset has been taken from the machine learning warehouse of University of California. Missing values of the instances are adjusted using mean value attribute method and the dimensions are trimmed down using PCA which capitulate the seven attributes including class attribute. It has been presented that the results obtained by the algorithms in this paper are better than the other techniques of the compared research papers.

Keywords*: Hepatitis C Virus (HCV), Data Mining, Clustering, Classification, Naïve Bayesian Classification, Decision Tree and Fuzzy C Mean (FCM).*

1. Introduction

Hepatitis is an inflammation of the liver based on different aetiologies. Clinician distinguished acute from chronic hepatitis [1]. Liver is the body's largest single organ and is necessary for life, Hepatitis causes the swelling or redness of the liver and characterized by the presence of inflammatory cells in the tissue of the organ. The condition can be self-limiting or can progress to fibrosis (scarring) and cirrhosis [2].

Five viruses have been identified that can primarily manifest clinically as acute hepatitis: hepatitis A virus (HAV), hepatitis B virus (HBV), hepatitis C virus (HCV), hepatitis D virus (HDV) and hepatitis E virus (HEV); While HAV and HEV are transmitted by the faecal-oral route and are often associated with acute icteric hepatitis, they do not lead to chronic infection. By contrast, HCV, HBV and HDV are transmitted parenterally and sexually, they are the most common cause of human viral infections leading to chronic hepatitis [1].

1.1. Introduction to Hepatitis C

HCV can be found in blood and possibly in many other body organs, but its favorite hideout is the liver. As the body repeatedly attempts to destroy the virus in the liver, inflammation of the liver occurs. Most people who have been infected with HCV do not have clinically recognized episode of acute hepatitis, but still go on to develop chronic hepatitis C [3].

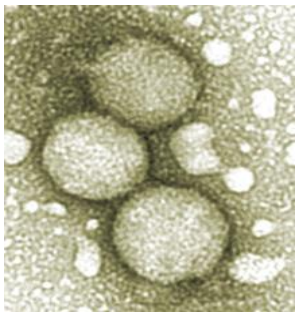

Figure 1: Image of particles isolated from HCV

Acute hepatitis C is a short-term infection with the hepatitis C virus. Symptoms can last up to 6 months. The infection sometimes clears up because your body is able to fight off the infection and get rid of the virus. Chronic hepatitis C is a long-lasting infection with the hepatitis C virus. Chronic hepatitis C occurs when the body can't get rid of the hepatitis C virus. Most hepatitis C infections become chronic. Without treatment, chronic hepatitis C can cause scarring of the liver, called cirrhosis; liver cancer and liver failure. Chronic hepatitis C is treated with drugs that slow or stop the virus from damaging the liver. A liver transplant may be necessary if chronic hepatitis C causes liver failure, Drug treatment often must continue because hepatitis C usually comes back after liver transplantation [3] [4] [5].

There are many studies carried on Hepatitis patients' data which really improved the diagnosis, safety and other measure of Hepatitis patients. Kedziora et al [6] demonstrated that phylogenetic trees and Hamming distances best reflect the differences between virus populations present in the organisms of patients who responded positively and negatively to the applied therapy. Tahseen A. Jilani [7] presented an automatic diagnosis system based on Neural Network for hepatitis virus which deals with the mixture of feature extraction and classification. Kemal Polat, SalihGüne [8] presented a novel method for diagnosis of hepatitis disease which is based on a hybrid method that uses feature selection (FS) and artificial immune recognition system (AIRS) with fuzzy resource allocation mechanism. Tahseen A. Jilani 2011 [9] scrutinized the factors that dole out significantly to augmenting the risk of hepatitis-C virus. Avendao [10] formulated a model to describe the dynamics of hepatitis C virus (HCV) considering four populations: uninfected liver cells, infected liver cells, HCV and T cells. Analysis of the model reveals the existence of two equilibrium states, the uninfected state in which no virus is present and an endemically infected state, in which virus and infected cells are present. There exists a threshold condition that determines the existence and stability of the endemic equilibrium.

Thair Nu Phyu [11] provide a comprehensive review of different classification techniques including decision tree induction, Bayesian networks, k-nearest neighbour classifier, case-based reasoning, genetic algorithm and fuzzy logic techniques. M. H. Marghny, Rasha M. Abd El-Aziz and Ahmed I. Taloba [12] proposed a technique to handle large scale data, which can select initial clustering center purposefully using Genetic algorithms (GAs), reduce the sensitivity to isolated point, avoid dissevering big cluster, and overcome deflexion of data in some degree that caused by the disproportion in data partitioning owing to adoption of multi-sampling. T.Karthikeyan and P.Thangaraju [13] provided the study on various classification algorithms namely, Bayes, NaiveBayes, Bayes.BayesNet, Bayes. NaiveBayesUpdatable, J48, Randomforest, and Multi Layer Perceptron. It analyzes the hepatitis patients from the UC Irvine machine learning repository. The results of the classification model are accuracy and time. Finally, it concludes that the Naive Bayes performance is better than other classification techniques for hepatitis patients.

1.2. Data set description

The data of Hepatitis C patients' has been taken from University of California repository [14]. The data contains 20 attributes (including a class attribute) and 155 instance out of which 32 belongs to death cases and rest of them belongs to live cases. Since there are missing values in data which has been filled using the technique mean value attribute [15]; and as the dimension of a huge data set can be trimmed down by using Principal Component Analysis which is considered as one of the most prevalent and useful statistical method [7], therefore after applying PCA the dimensions are trimmed to seven attributes (including a class attribute), Table 1 provides the description of these attributes.

1.3. Dividing data into training and test data

The dataset contains 155 instances, which is divided into training and test data. For this purpose we separated the dataset into life and death cases, and then generated a random number and taken 60% data from life cases and 60% data from death cases which has been considered as training set while remaining 40% of both cases are considered as test data sets.

2. Naïve Bayesian Classification

Naïve Bayesian is one of the most efficient and effective supervised learning algorithm which is based on Bayes' theorem. Naïve Bayesian helps to find the uncertainty of the model (new instance) in a principled way by determining the probabilities of the outcomes (training set). Naïve Bayesian

classification has comparable performance with decision tree and selected neural network classification techniques; it also provides high accuracy and speed when applied to large database [16].

$$\boldsymbol{P(H|X) = \frac{P(X|H)P(H)}{P(X)}}$$

Where:
X: be a data sample whose class label is unknown and P(X) is its probability.
H: a hypothesis that X belongs to class C and P (H) is its probability.
P(X | H) is the probability of X conditioned on H.

Naïve Bayesian classification uses distinct approaches for categorical and continuous valued attribute. Since in this paper the data contains continuous valued attribute therefore we discuss about Gaussian distribution used by the Naïve Bayesian algorithm, if data have continuous valued attribute. Beside this the technique has been applied in two ways, 1st it is applied on the existing data set and then it is applied on transformed data. Table 3 shows the accuracy of both cases.

2.1. Algorithm applied on Hepatitis C patients' data

Step 1: Find the Class Probabilities $P(C_1)$ on training set.

$$\mathbf{P(C_i) = \frac{\sum instances\ \epsilon\ C_i}{\sum instances}}$$

Step 2: Find the mean (μ) and standard deviation ($\boldsymbol{\sigma}$) of every attributes except the class attribute on training set.

$$\mu = \frac{1}{n}\sum_{i=1}^{n}(X_i)\ \boldsymbol{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu)^2}$$

Step 3: Find the conditional independence on test data using Gaussian distribution.

$$\mathbf{P(x_k|C_i) = g(x_{k_i}, \mu_{C_i}, \sigma_{C_i}) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}}$$

Where $\mathbf{x_k}$ refers to the values of attribute for instances x; $\boldsymbol{\mu_{C_i}}$ and $\boldsymbol{\sigma_{C_i}}$ are the mean and standard deviation respectively of the values of attribute for training instances of class $\mathbf{C_i}$.

Step 4: Calculate the Posterior Probabilities

$$\mathbf{P(C_i|X) = P(X|C_i) * P(C_i)}$$

Step 5: The classifier will predict that x belongs to the class $\mathbf{C_i}$ if and only if

$$\mathbf{P(C_i|X) > P(C_j|X)}$$

Where $\mathbf{1 \le j \le m, j \ne 1}$; m is the number of classes

Step 6: Validate the Results using

$$\mathbf{Accuracy = \left(\frac{TruePositive + TrueNegative}{Total}\right) * 100}$$

2.2. Applying transformation on data sets

In order to improve the results we have applied transformation technique on attribute level. For this the attribute Bilirubin has been transformed with Logistic regression which is one of the methods for describing the relationship between a categorical response variable and a set of predictor variables [17]. The other attribute like Alk Phosphate, Sgot, Albumin, Protime are transformed by taking the log of the values. After applying the transformation, transformed attribute has been scaled through multiplying the attributes by 100 and then Naïve Bayesian Classification has been applied as described earlier.

Table 3: Cases and accuracy

Cases	Accuracy of the result
Without Applying Any Treatment to the Data Set	95.102%
After Applying Transformation and Scaling	96.7347%

3. Decision Tree

Decision tree programs construct a decision tree T from a set of training cases. In order to define information gain precisely; we need to define a measure commonly used in information theory, called entropy that characterizes the impurity of an arbitrary collection of examples

$$\text{Entropy}(S) = - pp^* \log_2 (pp) - pn^* \log_2 (pn)$$

Where pp is the proportion of positive examples in S and pn is the proportion of negative examples in S [18].

Given entropy as a measure of the impurity in a collection of training examples, we can now define a measure of the effectiveness of an attribute in classifying the training data called information gain, the information gain, Gain (S, A) of an attribute A, relative to a collection of examples S, is defined as

$$\mathbf{Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|Sv|}{|S|} Entropy\ (Sv)}$$

where Values(A) is the set of all possible values for attribute A, and Sv is the subset of S for which attribute A has value v (i.e., Sv = {s ÎS | A(s) = v}). Gain (S,A) is the expected reduction in entropy caused by knowing the value of attribute A.

The process of selecting a new attribute and partitioning the training examples is repeated for each non-terminal descendant node, this time using only the training examples associated with that node. Attributes that have been incorporated higher in the tree are excluded, so that any given attribute can appear at most once along any path through the tree. This process continues for each new leaf node until either every attribute has already been included along this path through the tree, or the training examples associated with this leaf node all have the same target attribute value (i.e., their entropy is zero).

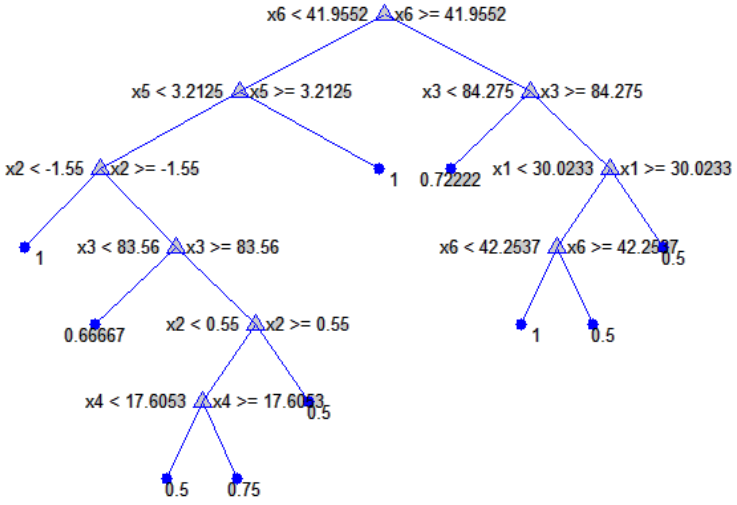

Figure 2: Decision Tree

We have applied MIN-MAX and Difference transformations on our data set which provides the accuracy of 96% accurate result. Min-max normalization is used to transform the data values for number attribute into the range [0, 1] and in Difference transformations each category of the predictor variable except the first category is compared to the average effect of previous categories also known as reverse Helmert contrasts [19].

Table 4 shows the accuracy obtained by the Decision Tree technique after executing the algorithm more than twenty numbers of time.

Table 4: Result obtained

CORRECTLY CLASSIFIED INSTANCES	INCORRECTLY CLASSIFIED INSTANCES	ACCURACY (%)
90	3	96

4. Fuzzy c-mean Clustering (FCM)

Clustering is the task of assigning a set of objects into groups called clusters, so that the objects in the same cluster are more similar to each other than to those in other clusters. Clustering is a main task of explorative data mining, and a common technique for statistical data analysis used in many fields, including learning, pattern, image analysis, information retrieval, and bioinformatics [20].

Fuzzy c-means is a method of clustering which allows one piece of data to belong to two or more clusters. This method is frequently used in pattern recognition. Any point x has a set of coefficients giving the degree of being in the kth cluster$\mathcal{W}_{\kappa}(x)$. With fuzzy c-means, the centroid of a cluster is the mean of all points, weighted by their degree of belonging to the cluster.

$$C_{\kappa} = \frac{\sum_x \mathcal{W}_{\kappa}(x)\, x}{\sum_x \mathcal{W}_{\kappa}(x)}$$

The degree of belonging$\mathcal{W}_{\kappa}(x)$ is related inversely to the distance from x to the cluster center as calculated on the previous pass. It also depends on a parameter m that controls how much weight is given to the closest centre

Fuzzy c-means (FCM) is based on minimization of the following objective function

$$\mathbf{J_m} = \sum_{\mathbf{i=1}}^{\mathbf{N}} \sum_{\mathbf{j=1}}^{\mathbf{C}} \mathbf{u_{ij}^m}\, ||\mathbf{x_i} - \mathbf{c_j}||$$

where m is any real number greater than 1, uij is the degree of membership of xi in the cluster j, xi is the ith dimensional measured data, cj is the d-dimension center of the cluster, and ||*|| is any norm expressing the similaritybetween any measured data and the center. Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership uij and the cluster centers cj by:

$$\mathbf{u_{ij}} = \frac{\mathbf{1}}{\sum_{\mathbf{k=1}}^{\mathbf{c}} \left(\frac{\left|\left|\mathbf{x_i} - \mathbf{c_j}\right|\right|}{\left|\left|\mathbf{x_i} - \mathbf{c_k}\right|\right|} \right)^{\frac{2}{\mathbf{m-1}}}} \qquad \mathbf{c_j} = \frac{\sum_{\mathbf{i=1}}^{\mathbf{N}} \mathbf{u_{ij}^m} \cdot \mathbf{x_i}}{\sum_{\mathbf{i=1}}^{\mathbf{N}} \mathbf{u_{ij}^m}}$$

This iteration will stop when$max_{ij} \left\{ \left| u_{ij}^{(k+1)} - u_{ij}^{(k)} \right| \right\} < \varepsilon$, where ε is a termination criterion between 0 and 1,

Whereas k is the iteration steps, this procedure converges to a local minimum or a saddle point of Jm [21].

The process of converting the fuzzy output is called defuzzification. Before an output is defuzzified all the fuzzy outputs of the system are aggregated with a union operator [22]. The union is the max of the set of given membership functions and can be expressed as:

$$\boldsymbol{\mu} = \cup_{\boldsymbol{i}} \left(\boldsymbol{\mu}(\boldsymbol{X})\right)$$

4.1. Algorithm applied on Hepatitis C patients' data

Step 1: Applying fuzzy c-mean Cluster requires two clusters because there exist two cases of life and death.

[center,U,obj_fcn] = fcm(data,cluster_n)

Step 2: Applying maximum deffuzification so we will put instances in different classes: if (U(1,i) > U(2,i)) Class = 1 else Class = 2

Step 3: Finally finding the accuracy using the following formula:

$$\textbf{Accuracy} = \left(\frac{\textbf{TruePositive} + \textbf{TrueNegative}}{\textbf{Total}}\right) * \textbf{100}$$

After repeating the iteration more than thirty times we obtained the average accuracy of 99.18%.

5. Results and Comparision

Before proceeding for model fitting, we have filled the missing values using mean value attribute technique. Then we reduce the dimension using principal component analysis as because the data have the problem of dimensionality's curse. After data reduction, the seven independent variables are Age, Bilirubin, Alk Phosphate, Sgot, Albumin and Protime.

Table 5 shows the accuracies of various data mining techniques applied on HCV patients' data. Table 6 shows the shows the accuracy percentage of classification and clustering technique applied on this paper.

TABLE 5: Accuracies obtained by using hepatitis diagnostic methods

Used Method	Article Author's	Accuracy (%)
ANN	Tahseen A. Jilani	89.6
RBF	Özyıldırım, Yıldırım, et al.	83.75
Naïve Bayes and semi NB	Stern and Dobnikar	86.3
15NN, stand. Euclidean	Grudzinski	89
FSM without rotations	Adamczak	88.5
IncNet	Norbert Jankowski	86
LVQ	Stern and Dobnikar	83.2
REGRESSION MODEL	Tahseen A. Jilani	89.6
CART (decision tree)	Stern and Dobnikar	82.7
RBF (Tooldiag)	Adamczak	79
Bayes.NaiveBayes	T.Karthikeyan, P.Thangaraju	84
Bayes.BayesNet	T.Karthikeyan, P.Thangaraju	81
Random forest	T.Karthikeyan, P.Thangaraju	83
J48	T.Karthikeyan, P.Thangaraju	83

LFC	Stern and Dobnikar	81.9
MLP	Özy ıldırım, Yıld ırım, et al.	74.37
ASR	Stern and Dobnikar	85
GRNN	Özy ıldırım, Yıld ırım, et al.	80
1NN	Stern and Dobnikar	85.3

TABLE 6: Proposed methods

ALGORITHM	ACCURACY (%)
NA ÏVE BAYESIAN CLASSIFICATION	95.102
NA ÏVE BAYESIAN CLASSIFICATION (On Transformed Data)	96.7347
DECISION TREE	96
FUZZY C-MEAN CLUSTERING (FCM)	99.1837

In proposed methods table 6 we found that the FCM have more accurate answer than other but as there are only two classes in the data set and clustering technique always give better accuracy for less number of classes, beside this the classification technique applied on this paper have better results than the methods used on other mentioned papers of table 5, we applied decision tree after transforming the data where the Na ïve Bayesian classification applied on non transformed and transformed data respectively. We have observed that the transformation improves the accuracy percentage of applied algorithm.

6. Conclusion and Future Studies

This paper provides the study on various data mining technique to investigate the factors of high prevalence of the risk of hepatitis C virus. A major challenge facing healthcare organizations (hospitals, medical centers) is the provision of quality services at affordable costs. Quality service implies diagnosing patients correctly and administering treatments that are effective. Poor clinical decisions can lead to disastrous consequences which are consequently unacceptable; therefore the focus is on using different algorithms for effective prediction of hepatitis C virus. The proposed work can be further enhanced and expanded for the automation of Hepatitis C virus prediction. Real data from Health care organizations and agencies needs to be collected and all the proposed techniques will be compared for the optimum accuracy. We also aim to implement K-mean, SVM and other data mining algorithm.

Acknowledgments

We want to give our special thanks to Badar Sami, Usman Amjad and Muhammad Hassan Ali for their support with the evaluation.

References

[1] Olaf Weber and Ulrike Protzer, Comparative Hepatitis, Springer Publisher.
[2] Information regarding hepatitis diseases available at: http://en.wikipedia.org/wiki/Hepatitis
[3] Beth Ann Petro Roybal, Hepatitis C: A personal Guide to Good Health, Third edition, Ulysses Press
[4] Information regarding Hepatitis C from National Digestive Diseases Information Clearinghouse (NDDIC), available at: http://digestive.niddk.nih.gov/ddiseases/pubs/hepc_ez/#1
[5] Information regarding Hepatitis C, available at:

http://www.british-liver-trust.org.uk/liver/hepatitis-c.html

[6] Kedziora P., Figlerowicz M., Formanowicz P., AlejskaM., Jackowiak P., Malinowska N., Fratczak A., Blazewicz J., and Figlerowicz M., "Comp utationalMethods in Diagnostics of Chronic Hepatitis C", Bulletin of the Polish Academy of Sciences, TechnicalSciences, 53 (3), 2005, pp.273-281

[7] Tahseen A. Jilani, Huda Yasin and Madiha Mohammad Yasin, "PCA-ANN for Classification of Hepatitis-C Patients", International Journal of Computer Applications, 14 (7), 2011.

[8] Polat K., Gunes S., "Hepatitis disease diagnosis using a new hybrid system based on feature selection (FS) and artificial immune recognition system with fuzzy resource allocation", Digital Sign al Processing 16, 2006, pp.889–901.

[9] Tahseen A. Jilani, Huda Yasin and Madiha Danish, "Hepatitis-C Classification using Data Mining Techniques", International Journal of Computer Applications, 24 (3), 2011.

[10] Avendao R., Esteva L., Flores J. A., et al., "A Mathematical Model for the Dynamics of Hepatitis C", Computational and Mathematical Methods in Medicine 4(2), 2002, pp.109 -118.

[11] Thair Nu Phyu, "Survey of Classification Techniques in DataMining", Proceedings of the International Multi Conference of Engineers and Computer Scientists, 2009 (I), IMECS 2009, Hong Kong.

[12] M. H. Marghny, Rasha M. Abd El-Aziz and Ahmed I. Taloba, "An Effective Evolutionary Clustering Algorithm: Hepatitis C case study", International Journal of Computer Applications, 34 (6), 2011.

[13] T.Karthikeyan and P.Thangaraju, "Analysis of Classification Algorithms Applied to Hepatitis Patients", International Journal of Computer Applications, 62 (15), 2013.

[14] Blake, C. L., & Merz, C. J. (1996), UCI repository of machine learning databases. Available at: http://archive.ics.uci.edu/ml/datasets/Hepatitis

[15] Edgar A. and Caroline R., "The treatment of missing values and its effect in the classifier accuracy", Department of Mathematics, University of Puerto Rico at Mayaguez.

[16] Jiawei Han, Micheline Kamber and Jian Pei, Data Mining Concepts and Techniques, Third Edition, Morgan Kaufmann Publishers.

[17] Daniel T. Larose, Data Mining Methods and Models, John Wiley and Sons, Inc.

[18] Information about decision tree available at http://dms.irb.hr/tutorial/tut_dtrees.php

[19] Rob Sullivan, Introduction to Data Mining for the Life Sciences, Springer Publisher.

[20] Information about clustering available at http://en.wikipedia.org/wiki/Cluster_analysis

[21] Jang J.-S.R., Sun C.-T and Mizutani E.(2003), Neuro-fuzzy computing and Soft computing Prentice Hall India.

[22] Klir G. J., and Yuan Bo (2005), Fuzzy Sets and Fuzzy Logic, Prentice Hall India.