*J. Appl. Environ. Biol. Sci.*, 4(9S)217-223, 2014



ISSN: 2090-4274
Journal of Applied Environmental and Biological Sciences
*www.textroad.com*

# A Comparative Study of Data Mining Techniques for Hcv Patients' Data

**Tahseen A. Jilani[1], Muhammad Shoaib[2], Rehan Rasheed[3] and Bilal Ur Rehman[4]**

Department of Computer Science, University of Karachi, Pakistan

*Received: September 12, 2014*
*Accepted: November 23, 2014*

## 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. Machine learning warehouse of University of California is the source from which the dataset has been obtained. Missing values 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. The liver get swelled or redish and characterized by the occurrence of inflammatory cells in the tissue of organ due to Hepapatis C. The condition can be self-limiting or can steps forward to fibrosis (scarring) and cirrhosis [2].

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 frequently attempts to demolish the virus in the liver due to which 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].

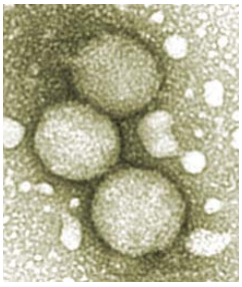

**Figure 1. Image of particles isolated from HCV**

There are many studies carried on Hepatitis patients' data which really improved the diagnosis, safety and other measure of Hepatitis patients. Kedziora et al [4] identified the reflection between virus population available in the organism of patients using phulogenetic trees and Hamming distances. Tahseen A. Jilani [5] implemented the feture extraction and proposed a mechanism based on Neural Network for the diagnosis of hepatitis virus. Kemal Polat, Salih Güne [6] presented a new technique to diagnose hepatitis disease, using fuzzy resource allocation mechanism with artificial immune recognition system and feature selection. Tahseen A. Jilani 2011 [7] examined the aspect that dole out extensively to enhance the hazard of hepatitis-C virus.

*** Corresponding Author:** Dr. Tahseen A. Jilani, Department of Computer Science, University of Karachi, Pakistan. Email: tahseenjilani@uok.edu.pk

Thair Nu Phyu [8] presented a comprehensive analysis of different classification method a including Bayesian networks, case-based reasoning, decision tree induction, genetic algorithm and fuzzy logic techniques k-nearest neighbour classifier. T.Karthikeyan and P.Thangaraju [9] provided the study on various classification algorithms namely, Bayes, Naïve Bayes, Bayes. Bayes Net, Bayes. Naïve Bayes Updatable, J48, Randomforest, and Multi Layer Perceptron. From the UC Irvine machine learning repository, It examines the hepatitis patients. Accuracy and time were the results of classification model. It has been concluded that for hepatitis patients the Naive Bayes performance is superior than other classification technique.

1.2. **Data set description.** The data of Hepatitis C patients' has been taken from University of California repository [10]. 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 [11]; and the dimensions has been reduced by using Principal Component Analysis [5], 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 for taking 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.

**Table 1. Description of attributes selected after data preparation and data reduction step**

| Attribute Name | Type | Descrption |
|---|---|---|
| **Class** | Live or Death | Die/ Live |
| **AGE** | Scaled Attribute | In the dataset, smallest age is 7 and largest age is 78. |
| **BILIRUBIN** | Continuous Attribute | A pigment fundamentally derived from the go down of haemoglobin from red blood cells damaged in the spleen |
| **ALK PHOSPHATE** | Scaled Attribute | It is an enzyme exists in the blood. |
| **SGOT** | Scaled Attribute | Serum glutamic oxaloacetic transaminase is an enzyme generally exists in serum. Also, it is present in heart. |
| **ALBUMIN** | Continuous Attribute | Serum albumin is the chief protein of blood plasma as well as of other serous solutions. |
| **PROTIME** | Scaled Attribute | Prothrombin is a predecessor of thrombin which is produced in the liver. |

## 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 the determination of outcomes' probabilities (training set). Also it is accurate and fast on the application to huge database [12].

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \quad (1)$$

Where:

X: be a data sample and P(X) is its probability for identifying the class label.

H: a hypothesis for identifying the belonging of X in class C and P (H) is its probability.

P(X | H) is the conditional probability of X given H.

Naïve Bayesian classification use different approaches for categorical and continuous valued attribute. Since in this paper the data contains continuous valued attribute therefore we have discuss about Gaussian distribution used by the Naïve Bayesian algorithm, 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.

$$P(C_1) = \frac{\sum \text{instances } \epsilon\, C_1}{\sum \text{instances}} \quad (2)$$

Step 2: Find the mean (μ) and standard deviation (**σ**) of every attributes except the class attribute on training set.

$$\mu = \frac{1}{n}\sum_{i=1}^{n}(X_i) \qquad \sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu)^2} \tag{3}$$

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

$$P(x_k|C_i) = g(x_{k_i}, \mu_{C_i}, \sigma_{C_i}) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{4}$$

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

Step 4: Calculate the Posterior Probabilities

$$\mathbf{P(C_i \mid X) = P(X \mid C_i) * P(C_i)} \tag{5}$$

Step 5: The classifier will predict that x $\epsilon$ $C_i$ iff

$$\mathbf{P(C_i \mid X) > P(C_j \mid X)} \tag{6}$$

Where $1 \leq j \leq m$, $j \neq 1$;; m is the number of classes

Step 6: Validate the Results using

$$\text{Accuracy} = \left(\frac{\text{TruePositive} + \text{TrueNegative}}{\text{Total}}\right) * 100 \tag{7}$$

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 method to describe the association among a categorical response variable and a set of predictor variables [13]. The other attribute like Alk Phosphate, Sgot, Albumin, Protime are transformed by taking the log of values. After applying the transformation, attributes have been scaled through multiplying the attributes by 100 and then Naïve Bayesian Classification has been applied as describe earlier.

2.3. **Accuracy of the result**

**Table 2. Cases and accuracy of Naïve Bayesian Implementation**

| 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

These programs used a set of training cases in the construction of a decision tree T. to exactly describe information gain; first a measure is to be defined normally used in information theory, termed as entropy which measures the disorderness in the information that describes the impurity of an arbitrary collection of examples.

**Entropy(S) = - pplog2 pp – pnlog2 pn** (8)

Here pp is the ratio of constructive examples in S and pn is the proportion of pessimistic examples in S [14]. The information gain, Gain (S, A) of a feature A, relative to set of examples S, is defined as

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{v \in \text{Values}(A)} \frac{|Sv|}{|S|} \text{Entropy}(Sv) \tag{9}$$

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 Î S | A(s) = v}).

Gain (S, A) is the expected caused of reduction in entropy by finding the value of feature A. The method of choosing a new feature and categorization the prepared examples is recurring for each non-terminal successor knob, this time we use only the prepared examples associated with that knob. Features

that have been included higher in the tree are disqualified, so that every feature can show at most once along any pathway through the tree. This method repeat for each new knob until either every feature has already been incorporated along this pathway through the tree, or the prepared examples associated with this knob, all have the same target feature value (i.e., their disorderness is reduced to zero).

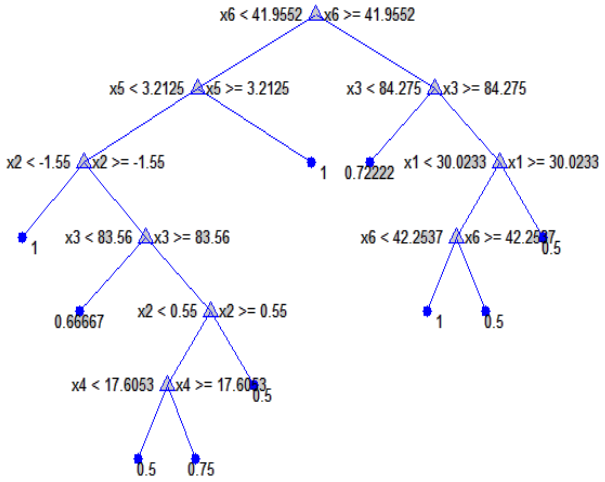


**Figure 2. Decision Tree**

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 [15].

Table 3 shows the accuracy obtained by the Decision Tree technique after executing the algorithm more than twenty numbers of time.

**Table 3. Result obtained**

| Correctly classified instances | Incorrectly classified instances | Accuracy (%) |
|---|---|---|
| **90** | 3 | 96 |

## 4. FUZZY C-MEAN CLUSTERING (FCM).

Fuzzy c-means is one of the technique of clustering which permit one part of data to fit in to two or more clusters. The existence in the kth cluster $w_K$ (x) is defined by the set of coefficients of any point x. With fuzzy c-means, depending on the level of belonging to the clusters the centroid of cluster is measured as the mean value of all points.

$$C_\kappa = \frac{\sum_x W_\kappa(x)\, x}{\sum_x W_\kappa(x)} \tag{10}$$

The rank of fitting into $w_K$ (x) is inversely correlated to the distance to the cluster center from x as previously defined. Which depends also on parameter m that controls the level of weight given to the closest centre. The mininmization of following objective function on which it is based upon.

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m}\, ||x_i - c_j||^2 \tag{11}$$

where m >1 belongs to real number, $u_{ij}$ is the degree of membership of $x_i$ in the cluster j, $x_i$ is the measurement of datt in ith dimensional, $c_j$ defines the d-dimension center of cluster, and ||*|| is any norm showing the likeness among the center and any calculated data. This function leads to Fuzzy partitioning after moving through an iterative optimization, with renewing of membership $u_{ij}$ and the cluster centers $c_j$ by:

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

(12)

$max_{ij}\{|u_{ij}^{(k+1)} - u_{ij}^{(k)}|\} < \varepsilon$

is the criteria to stop the iteration, Where ∈a termination criterion between 0 and 1, k is the iteration steps, this process meets to a local minimum or a saddle point of Jm [16].

**Defuzzification** is the process of alteration of fuzzy output.Before applying defuzification all the fuzzy outputs of the system are aggregated using a union operation [17], which is the max for the set of given membership functions and can be calculated using :

$$\mu = \cup_i (\mu(X)) \quad (13)$$

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)** (14)

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** (15)

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

$$\text{Accuracy} = \left(\frac{\text{TruePositive} + \text{TrueNegative}}{\text{Total}}\right) * 100 \quad (16)$$

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

**TABLE 4. Accuracies obtained by using hepatitis C 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 |
| **MLP** | Özy ıldırım, Yıld ırım, et al. | 74.37 |
| **GRNN** | Özy ıldırım, Yıld ırım, et al. | 80 |
| **1NN** | Stern and Dobnikar | 85.3 |

**TABLE 5. Proposed methods**

| Algorithm | Accuracy (%) |
|---|---|
| **NAÏVE BAYESIAN CLASSIFICATION** | 95.102 |
| **NAÏVE BAYESIAN CLASSIFICATION (On Transformed Data)** | 96.7347 |
| **DECISION TREE** | 96 |
| **FUZZY C-MEAN CLUSTERING (FCM)** | 99.1837 |

In proposed methods table 5 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 4, we applied decision tree after transforming the data where the Naïve 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. In healthcare organizations providing the outstanding services involved diagnosing patients appropriately and managing the treatments in a valuable manner is becoming a core challenge. Impropar clinical decision can cause terrible results which are consequently unacceptable; therefore the focus is on using different algorithms for effective prediction of hepatitis C virus. This work can be more extended for the automation of Hepatitis C virus forecast. All the proposed techniques will be applied on real data from health care organizations and agencies and optimum accuracy will be compared. 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, Muhammad Najamuddin 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. 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, Technical Sciences, 53 (3), 2005, pp.273-281.
5. 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.
6. 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.
7. Tahseen A. Jilani, Huda Yasin and Madiha Danish, "Hepatitis-C Classification using Data Mining Techniques", International Journal of Computer Applications, 24 (3), 2011.
8. 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.
9. T.Karthikeyan and P.Thangaraju, "Analysis of Classification Algorithms Applied to Hepatitis Patients", International Journal of Computer Applications, 62 (15), 2013.
10. Blake, C. L., & Merz, C. J. (1996), UCI repository of machine learning databases. Available at:

11. http://archive.ics.uci.edu/ml/datasets/Hepatitis
12. 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.
13. Jiawei Han, Micheline Kamber and Jian Pei, Data Mining Concepts and Techniques, Third Edition, Morgan Kaufmann Publishers.
14. Daniel T. Larose, Data Mining Methods and Models, John Wiley and Sons, Inc.
15. Information about decision tree available at http://dms.irb.hr/tutorial/tut_dtrees.php
16. Rob Sullivan, Introduction to Data Mining for the Life Sciences, Springer Publisher.
17. Fuzzy C-Mean Algorithm Info available at
18. http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/cmeans.html
19. Defuzzification information available at
20. http://enpub.fulton.asu.edu/PowerZone/FuzzyLogic/chapter%206/frame6.html