IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

Analysis of HCV Risk Factor among PWID's in India- An Approach using Naive Bayes Classifier

M. Gomathy[1], Saravanamurthy P. Sakthivel[2]

Assistant Professor, Department of Computer Application, NBGSM College, NCR Delhi, Haryana, India[1]

Public Health Consultant[2]

Abstract: The huge amount of data generated by healthcare transactions are too complex and voluminous to be processed and analyzed by traditional methods as the way in which healthcare is financed is critical for equity in access to healthcare. At present the proportion of public resources committed to healthcare in India is one of the lowest in the world, with less than one-fifth of health expenditure being publicly financed. To overcome this issue the researchers, use data mining techniques. Data mining through various algorithms provides the methodology and technology to transform large amount of data into useful information for decision making and patients receive better, more affordable healthcare services. In this paper, Naive Bayes algorithm is used to predict the risk factors associated with HCV infection among People Who Inject Drugs (PWID's) in India. Naive Bayes algorithm are the most popular algorithms for rule based classification as it requires minimal number of attributes.

Keywords: Data Mining, Naive Bayes Algorithm, PWID, HCV.

I. INTRODUCTION

From the year of identification, 1989, [1,2] of Hepatitis C Virus (HCV) and the availability of analyzing the antibody to HCV, the epidemiology of HCV infection has been investigated in many populations. The major group infected and at risk of continuing infection with HCV in India, as in other countries, are people who inject drugs (PWID). This increased risk is associated with multiple factors, including the practice of sharing injecting equipment; especially needles and syringes, similar to Hepatitis B Virus (HBV) and Human Immunodeficiency virus (HIV). [3]

Data mining holds great potential for the healthcare industry to enable health systems to systematically use data and analytics to identify inefficiencies and best practices that improve care and reduce costs. In the present study, Naive Bayes Algorithm is used to consider the vital attributes necessary for deriving the prediction of risk factor associated with HCV infection among PWID's in India. The Bayesian Classification represents a supervised learning method as well as a statistical method for classification which also helps us to identify and solve diagnostic and predictive problems.

The data used in this article is from Integrated Biological and Behavioural Assessment (IBBA),2009. [4]. This survey captured Hepatitis C virus associated risk behaviours' among PWIDs in Dimapur of Nagaland in India. There are 440 records collected from this district of Nagaland among PWID's. Each record has the same structure, consisting of a number of attributes or value pairs. One of these attributes represents the category of the record which says the prevalence of high risk of HCV or low risk of HCV from the collected information. The category attributes are assumed to take the values {high risk, low risk}. The non-category data which has been chosen among various attributes have association on the risk factor of HCV. The domain values which specify the risk factors among the PWID's are defined for the present investigation includes, (i) Marital status (MStatus), (ii) Type of drug(TDrug), (iii) Location of injecting practice (LOC), (iv) Shared Needles/Syringe(SharN/S), (v) No.of persons shared(No.Shar)., (vi) Frequency of Needle/Syringe sharing (Freq.Shar), (vii) Injecting from Pre filled Syringe(Inj Pre Syn), (viii) Sharing of Common Container(Shar CC).

II. LITERATURE REVIEW

Many studies had predicted risk factors associated with HCV infection among PWIDs globally including India from clinical information using different data mining techniques. Some of them are discussed here.

Bendi Venkata Ramana et.al. utilized Classification Algorithms for evaluating the classification performance in terms of Accuracy, Precision, Sensitivity and Specificity in classifying liver patients dataset.and proved KNN, Back propagation and SVM are giving better results. [5]

M.Gomathy and Dr Vani Perumal utilized c4.5 algorithm to rank the risk factors associated with HCV among PWID's in India and also used decision tree for effective decision making.[6]

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

Chaitrali S. Dangare et.al has structured prediction systems for Heart disease using more number of input attributes. They used the data mining classification techniques like Decision Trees, Naive Bayes and Neural Networks. The performances of these techniques are compared based on accuracy. [7]

Dipali Bhosale et.al used Naïve Bayes algorithm for feature selection. Co-relation based Feature Selection, Wrapper, and Information Gain WERE USED on the data sets. Then, by using these three feature selection techniques they separate feature subsets and derived the final results [8].

Dr. S. Vijayarani et.al used Naïve Bayes Algorithm and Support Vector Machine for the prediction of Liver disease. They predicted normal liver diseases, CBCL, Acute Hepatitis and Outliers using six attributes [9].

Ms.Ankita et.al utilized Naïve Bayes Classifier for the prediction of Swine Flu disease. They have used the values of eight different attributes for the prediction of the disease. [10]

Jyoti Soni et.al. applied various datamining techniques in prediction of heart diseases. The author compared the performance of predictive data mining technique on the same dataset proved that Decision Tree outperforms and Bayesian classification is having similar accuracy in the prediction of diseases. They used genetic algorithm in Bayesian classification which reduced the actual data size and got the optimal subset of attribute for the prediction of heart disease. [11]

K.Srinivas et al. examined the potential use of classification based data mining techniques such as Rule based, Decision tree, Naïve Bayes and Artificial Neural Network to massive volume of healthcare data. For data preprocessing and effective decision making One Dependency Augmented Naïve Bayes classifier (ODANB) and naive credal classifier 2 (NCC2) are used which increase the robust classifications when dealing with small or incomplete data sets. They discovered all the hidden patterns, their relationships between medical factors such as age, sex, blood pressure and blood sugar related to heart disease are established using data mining. [12]

JS Sartakhti et.al.used a novel machine learning method that hybridizes support vector machine (SVM) and simulated annealing (SA) techniques of data mining in diagnosis of hepatitis disease. The data is from hepatitis disease dataset from UCI repository. On applying the above method they obtained classification of 96.25%.The author has proved that the data mining technique, SVM-SA method can assist in the diagnosis of hepatitis. [13]

FM Ba-Alwi et.al. here used 7 different types of data mining algorithm namely, Naive Bayes, Naive Bayes updatable, FT Tree, KStar, J48, LMT, and Neural network for analyzing Hepatitis prognostic data. The results of the classification are accuracy and time. The study concludes that the Naive Bayes classification performance is better than other classification techniques for hepatitis dataset. [14]

KC Tan et.al. used hybrid evolutionary algorithm which identifies the appropriate attributes that would reduce the size of the dataset and allow more comprehensible analysis to extract patterns or rules. The author utilized two conventional machine learning algorithms for selecting the attributes. They used Genetic algorithms (GAs) which searches for the best attribute and Support Vector Machines (SVMs) classifies the patterns in the reduced datasets and are integrated effectively based on a wrapper approach. They also demonstrated that the GA-SVM hybrid produces good classification accuracy and a higher level of consistency that is comparable to other data mining algorithms. [15]

III. APPLICATION OF NAÏVE BAYES CLASSIFIER IN INJECTING DRUG USERS (PWID's) DATA SET

In this section the theory behind the Naïve Bayes classification algorithm, the prediction methodology and experimental results with IBBA dataset among PWID is discussed in detail.

A.Naive Bayes Algorithm
It is based on the Bayesian theorem which is particularly suited when the dimensionality of the inputs is high. Parameter estimation for naive Bayes models uses the method of maximum likelihood. Naive Bayes classifiers is a machine learning algorithm, also called probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions between the features. This algorithm requires minimal number of attributes.

The entire Naïve Bayes classification algorithm can be implemented in three steps. They are given below:
Step 1: Each data sample is represented as an n dimensional feature vector, X = (x1, x2... xn). This depicts n measurements made on the sample from n different attributes like A1, A2 ... An respectively.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

Step 2: Assume that there are n classes, C1, C2...Cn in a data sample X, the classifier will predict that X belongs to the class having the highest posterior probability, which is conditioned as: if and only if: P(Ci/X)>P(Cj/X) for all 1< = j< = n and j!= i Thus P(Ci|X) is maximized. As mentioned above, thus the class Ci for which P(Ci|X) is maximized is referred as maximum posteriori hypothesis.
Step 3: As described above, P(X) is constant for all classes, only P(X|Ci) P(Ci) need be maximized. But for the class, prior probabilities are unknown. Hence it is assumed that the classes are equally likely, and so that P(X|Ci) P(Ci) is maximized. If they are not equally likely, P(X|Ci) is maximized. The above described algorithm is applied among the data to predict the factors responsible for the HCV infection among the PWID's.

B. The Proposed Methodology
As mentioned above the observed IBBA-PWID dataset contains Four hundred and forty records of different values for nine different non-categorical attributes (refer Table I). Using the data table that contains attributes and class of the attributes, the homogeneity (or heterogeneity) is measured based on the classes. If a table is pure or homogenous, it contains only a single class. If a data table contains several classes, then it says that the table is impure or heterogeneous.

C. Experimental Results on prediction of risk factors for HCV infection among PWID's by Naive Bayes

Table 1 is the result on application of the algorithm. This also gives the probability of the various attributes

TABLE I : Probability of the attributes for the prediction of HCV infection among PWID'S

Attribute	**Possible Value**	**Prediction of risk Factor**	
		High	**Low**
MStatus	0-Married , 1- Unmarried	30/50 20/50	65/390 325/390
TDrug	1-Heroin , 2 –Other Drug	50/50 0/50	42/390 348/390
Loc	1- Own house, >1 Other place	50/50 0/50	88/390 302/390
Shar N/S	1-yes, 0-no	50/50 0/50	140/390 230/390
No. Shared	> 2, < 2	50/50 0/50	229/390 161/390
Freq.Shar	1-Every time >1 Not always	50/50 0/50	0/390 390/390
Inj Pre Syn	1-Every time >1 not always	50/50 0/50	7/390 383/390
Shar CC	1-Every time , >1 not always	50/50 0/50	25/390 365/390

From Table I, all possible probabilities conditioned on the target attributes used for the prediction of HCV infection among the PWID's is computed using Navie Bayesian algorithm. The results are as follows
From the given data, we wish to identify the risk factor according to the condition which satisfy
X = (MStatus=0, TDrug=1, Loc>1, SharN/S=1, No.
Shared>2, freq.Shar N/S=1, Inj Pre syn=1, Shar

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

CC=1)
We need to maximize P(X|Ci)P(Ci), for i = 1, 2.
P(Ci), the priori probability of each class, can be estimated based on the data is
P(C1)=P(risk=high)= 50/440 and
P(C2)=P(risk=low)= 390/440

The probabilities P(X|Ci), for i = 1, 2, is obtained as follows
P (MStatus =married/risk=high) = 30/50 =0.6
P (TDrug=Heroine/risk=high) = 50/50 =1

P (Shar N/S =yes/risk=high) = 50/50 P (No. Shared>2/risk=high) = 50/50
P (Freq.Shar N/S =1/risk=high) = 50/50=1
P (Inj Pre Syn =1/risk=high) = 50/50=1
P (Shar CC =1/risk=high) = 50/50=1
P (Loc >1/risk=high) = 50/50
Using the above probabilities, we obtain
P(X/risk=high) = (3/5)*(5/5)* *(5/5) *(5/5) *(5/5)
*(5/5) *(5/5) *(5/5)
=0.6

Similarly the probabilities for risk=low is calculated as follows

P(MStatus =married/risk=low) = 65/390 =0.166
P(TDrug=Heroine/risk= low) = 42/390 =0.107
P(Shar N/S =yes/risk= low) = 140/390 = 0.358
P(No. Shared>2/risk= low) = 229/390 = 0.587
P(Freq.Shar N/S =1/risk= low) = 0/390 = 0
P(Inj Pre Syn =1/risk= low) = 7/390 = 0.0179
P(Shar CC =1/risk= low) = 25/390 = 0.0641
P(Loc >1/risk= low) = 88/390 = 0.225

Using the above probabilities, we obtain P(X/risk=low) =(65/390)*(42/390)*(140/390)*
(229/390) * (0/390)*(7/390)*
(25/390)*(88/390)
= 0
From the above calculation, the result obtained is P(X/risk=high) =0.6 and P(X/risk=low) =0. The maximum value among P(X|Ci) *P(Ci) predicts the risk factor associated with attributes among the PWID's. Thus, by implementing the Naïve Bayes classifier and the information derived by using the Classification algorithm for the above data set shows that the attributes utilized to identify the possibility of HCV infection among the PWID's can be predicted using Naïve Bayes Algorithm

V. CONCLUSION and FUTURE WORK

Classification is the major data mining technique which is primarily used in healthcare sectors for medical diagnosis and predicting diseases. The findings indicates, that Naive Bayes theorem was helpful in indicating the association of key factors with HCV. This would assist HIV prevention intervention to routinely check the factors associated with HCV from their intervention data with this simple analytical tool.

REFERENCES

[1] Zanetti, Alessandro Remo, Global surveillance and control of hepatitis C Report of a WHO Consultation organized in collaboration with the Viral Hepatitis Prevention Board, Antwerp, Belgium. J Viral Hepat. 1999;6:35–47. [PubMed]
[2] Centre for Diseases Control and Prevention. 2015. Hepatitis C: 25 years Since Discovery. https://www.cdc.gov/knowmorehepatitis/media/pdfs/hepc-timeline.pdf
[3] Samiran Panda et.al "Alarming epidemics of human immunodeficiency virus and hepatitis C virus among injection drug users in the northwestern bordering state of Punjab, India: prevalence and correlates" International Journal of STD & AIDS,2013, DOI: 10.1177/0956462413515659.
[4] ICMR-FHI,2009. Integrated Biological and Behavioural Assessment (IBBA),Round 2.Indian Council of Medical research-Family Health International.
[5] 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
[6] M.Gomathy and Dr Vani Perumal " Prediction of risk Factor for HCV infection among PWIDs-An Approach using C4.5 ","International Journal of Advanced research in Computer and communication Engineering"Vol 6,Issue 11,Nov 2017

[7] Chaitrali S. Dangare and Sulabha S. Apte, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques", International Journal of Computer Applications, vol. 47, pp. 44-48, Jun.2012.
[8]Dipali Bhosale and Roshani Ade, "Feature Selection based Classification using Naive Bayes, J48 and Support Vector Machine", International Journal of Computer Applications, vol.99, pp. 14-18, Aug. 2014.
[9] Dr. S. Vijayarani, Mr.S.Dhayanand, "Liver Disease Prediction using SVM and Naïve Bayes Algorithms", International Journal of Science, Engineering and Technology Research, vol.4, pp. 816-820, Apr. 2015.
[10] Ankita R. Borkar and Dr. Prashant R. Deshmukh "Naïve Bayes Classifier for Prediction of Swine Flu Disease", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 5, pp. 120-123, Apr. 2015
[11] Jyoti Soni,Ujma Ansari,Dipesh Sharma, Sunita Soni,
"Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", International Journal of Computer Applications (0975 – 8887) Volume 17– No.8, March 2011
[12] K.Srinivas, B.Kavihta Rani , A.Govrdhan ," Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks", IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 02, 2010, 250-255
[13] JS Sartakhti, MH Zangooei, K Mozafari ,"Hepatitis disease diagnosis using a novel hybrid method based on support vector machine and simulated annealing (SVM-SA)", Computer Methods and Programs in Biomedicine, Volume 108, Issue 2, November 2012, Pages 570-579
[14] FM Ba-Alwi, HM Hintaya ," Comparative Study for Analysis the Prognostic in Hepatitis Data: Data Mining Approach" ,International Journal of Scientific & Engineering Research, Vol 4, Issue 8, August-2013 680 ISSN 2229-5518 IJSER © 2013 http://www.ijser.org
[15]KC Tan, EJ Teoh, Q Yu, KC Goh ," A hybrid evolutionary algorithm for attribute selection in data mining ",Expert Systems with Applications,Volume 36, Issue 4, May 2009, Pages 8616-8630

BIOGRAPHIES

Mrs M.Gomathy received her M.C.A degree from the Presidency College, Madras University and M.Phil from Bharathidasan University. She is currently working as Assistant Professor in the Department of Computer Application, NBGSM College, NCR Delhi, Haryana, India.

P.S. Saravanamurthy, a public health researcher with 15 years of public health research and program experience in India. Recipient of Fogarty fellowship from Albert-Einstein College of Medicine and completed his doctoral degree in Applied Microbiology from Department of Microbiology, Dr. A.L.M. Post Graduate Institute of Basic Medical Sciences, University of Madras. He has authored/co-authored 10 peer-reviewed articles in national and international journals.