Volume 4, Issue 1, January 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Comparison of Data Models for Hepatitis Diagnosis – Data Mining Technique

S Pushpalatha
MCA Dept. & Kadi Sarvavishwa Vidyalaya
India

Dr. Jagdesh Pandya
Manager –BISAG Gandhinagar
India

Abstract - Hepatitis is a lever disease which affects majority of the population in all age group. It is the major challenge for many hospitals and public health care services for diagnosing hepatitis. Proper diagnosis and accurate prediction of the disease on time can save many patients. Data mining is an efficient tool to diagnose hepatitis from large dataset and to predict the severity of the disease. This paper reviews different data mining techniques which are used to diagnosis hepatitis disease and shows the performance of different data mining techniques which were implemented.

Keyword: Hepatitis, Data Mining, Diagnose, Public health care, prediction.

I INTRODUCTION

Inflammation of liver is the main characteristic of hepatitis. Hepatitis occurs with limited or no symptoms which may lead to jaundice, anorexia and malaise. Hepatitis is now one of the most important causes of chronic liver disease in the world, and millions of people are at risk for its complications. It is one of the most common infectious diseases, causing an estimated 1.5 million deaths worldwide each year [1].

Viral hepatitis is an inflammation and damage to hepatocytes in the liver caused by at least six different viruses. These viruses called A, B, C, D, E, and G that are also called HAV, HBV,HCV, HDV, HEV, and HGV respectively[2].

The hepatitis A commonly occurs in children also named as infectious hepatitis. The hepatitis A is hepatitis type appears because of the hepatitis A virus (HAV). This hepatitis A exists in the stools, feces or poop of infected individuals. The hepatitis B arises because of hepatitis B virus (HBV) also named as serum hepatitis. Infection transmitted percutaneously, sexually, and prenatally from infected body fluids, such as blood, saliva, semen, vaginal fluids, tears, and urine, a contaminated blood transfusion, shared contaminated needles or syringes for injecting drugs, sexual activity with an HBV-infected person, and transmission from HBV-infected mothers to their newborn babies. The hepatitis C is diffused by direct contact with an infected person's blood. This appears due to the hepatitis C virus (HCV). Infection is often asymptomatic, but once established, chronic infection can progress to scarring of the liver (fibrosis), and advanced scarring (cirrhosis) which is generally apparent after many years. The hepatitis D caused by hepatitis D virus (HDV). HDV is spread through contact with infected blood, dirty needles that have HDV on them and unprotected sex (not using a condom) with a person infected with HDV. Hepatitis D causes swelling of the liver. Hepatitis E is caused by the virus HEV. Hepatitis E by drinking water infected with the virus. It causes swelling of the liver, but no long-term damage. This paper is organized as follows. Section 2 specifies the overview of the related work. Section 3 specifies the concept of Data Mining techniques, section 4 discusses the problems in the previous research, section 5 discusses the proposed framework and finally section 6 discusses the conclusion.

II RELATED WORK

1. Yılmaz Kayaa et al. [3] implemented a new hybrid medical decision support system based on rough set (RS) and extreme learning machine (ELM) for the diagnosis of hepatitis disease. RS-ELM consists of two stages in the first stage the redundant features have been removed from the data set through RS approach. In the second stage the classification process has been implemented through ELM by using remaining features. Hepatitis data set was taken from UCI machine learning repository has been used to test the proposed hybrid model. A major part of the data set (48.3%) includes missing values. As removal of missing values from the data set leads to data loss, feature selection has been done in the first stage without deleting missing values. In the second stage, the classification process was performed through ELM after the removal of missing values from sub-featured data sets that were reduced in different dimensions. The classification accuracy was about 96.49% using RS-ELM model.

2. Javed Salimi Sartakhti et al. [4] presented a novel machine learning method using hybridized Support Vector machine and simulated annealing for hepatitis diagnosis. It is a stochastic method used for difficult optimization problems. Dataset used in this study from the UCI machine learning database. The classification accuracy is obtained via 10-fold cross validation and accuracy of this method is 96.25%.

3. Duygu et al. [5] proposed an intelligent hepatitis diagnosis system using Principle Component Analysis and Least Square Support Vector Machine Classifier (PCA–LSSVM). This intelligent hepatitis diagnosis system was separated into

two phases: (1) the feature extraction from hepatitis diseases database and feature reduction by PCA, (2) the classification by LSSVM classifier. Feature extraction is important for pattern reorganization, if feature not chosen well it should be reduced for obtaining original feature. 19 features of hepatitis dataset were reduced to 10 features using PCA. In second phase, these reduced features are given to inputs LSSVM classifier. LSSVM classifier has 2 parameters the width of Gaussian kernels σ and the regularization factor C. By adjusting the parameter values of σ between 0.1 and 25 and by adjusting the parameter values of C between 1 and 100000 suitable for SVM predicate performances. 10 combinations of C and σ values were obtained and the best classification accuracy obtained was 96.12% from σ as 0.8 and C as 100.

4. G. Sathya Devi [6] proposed the application of CART algorithm in Hepatitis Disease Diagnosis using decision trees C4.5 algorithm, ID3 algorithm and CART algorithms. It classifies the hepatitis diseases and compares the effectiveness, correction rate among them. From that the CART derived model showed the extended definition for identifying (diagnosing) hepatitis disease provided a good classification accuracy of 83.2%.

5. A.H.Roslina et al. [7] Implemented a prediction of hepatitis prognosis using Support Vector Machines and Wrapper Method. To remove the noise features wrapper methods were used before classification. Support Vector Machines showed the accuracy in enforcing feature selection first. Features selection were implemented to minimize noisy or irrelevance data. The accuracy rate was increased concurrently the clinical lab test cost and time was reduced. This was achieved by combining Wrappers Method and SVM techniques.

6. Fadl Mutaher et al. [8] presented the comparative analysis in the prognostic of hepatitis data using Rough set technique over Multi- layer Neural Network using back-propagation algorithm. The prediction of the outcome is more specific and accurate using Rough set technique. Performance and time taken to run the hepatitis data is fast in Naive Bayes algorithm. The results obtained were compared with other algorithms like, Naive Bayes up-datable algorithm, FT Tree algorithm, Kstar algorithm, J48 algorithm, LMT algorithm and neural network. Attributes were fully classified and the result obtained was of 96.52%. Based on the experimental results the classification accuracy is found to be better using Naïve Bayes algorithm compared to other algorithms.

TABLE 1
COMPARISON OF LITERATURE REVIEWS

S.No	Reseacher	Publication	Technique	Performances
1	Yilmaz Kaya	Elsevier - 2013	RS-ELM [Rough set - Extreme Learning Machine]	The classification accuracy was 96.49% using RS-ELM model.
2	Javad Salimi Sartakhti	Elsevier 2011	Using support vector machine (SVM) and simulated annealing (SA).	Obtained classification accuracy of 96.25%
3	Duygu et. Al.	Elsevier - 2011	Compared with LS-SVM classifiers and PCA-LSSVM	Accuracy is 96.12%
4	G. SathyaDevi	IEEE 2011	Use of decision tree C4.5 algorithm, ID3 algorithm and CART algorithm	CART should the accuracy rate of 83.2%
5	A.H.Roslina	IEEE 2010	SVM and Wrapper method to remove noise feature before classification. [Used without feature selection & with feature selection]	Accuracy rate is 74.55%
6	Fadl Mutaher	IJSER 2013	Using Classification algorithms like Naive Bayies, FT Tree, KStar, J4.8	Naive Bayes showed the accuracy of 96.52%

III DATA MINING TECHNIQUES

A. Classification:

Classification is used to classify data into predefined class labels. Class in classification, is the attribute or feature in a data set, in which users are most interested. It is defined as the dependent variable in statistics. To classify data, a classification algorithm creates a classification model consisting of classification rules. Classification can be used to diagnose hepatitis and prognosis based on symptoms and health conditions [9]. In this there are two steps process consisting of training and testing. The first step is training which used to builds a classification model by analyzing training data containing class labels. The second step is testing. It examines a classifier using testing data for accuracy in which the test data contains the class labels or its ability to classify unknown objects for prediction. There are many classification algorithms like Naive Bayes, FT Tree, KStar, J48, Neural network.

B. Support Vector Machine

Support vector can be used for pattern classification [9] which has multilayer perceptrons and radial-basis function networks. An idea that is central to the construction of the support vector learning algorithm is the inner-product kernel between a support vector and the vector drawn form the input space. The support vectors are made up of small subset of the training data extracted by the algorithm. Support vector learning algorithms may be used to construct three types of learning machines like Polynomial learning machines, Radial-basis function networks, Two-Layer perceptrons.

C. Naive Bayesian

A Naive Bayesian classifier is a probabilistic statistical classifier. The term "naive" refer to a conditional independence among features or attributes. The "naive" assumption reduces computation complexity to a simple multiplication of probabilities. One main advantage of the Naive Bayesian classifier is its rapidity of use. That's because it is the simplest algorithm among classification algorithms [8]. Because of this simplicity, it can readily handle a data set with many attributes. In addition, the naive Bayesian classifier needs only small set of training data to develop accurate parameter estimations because it requires only the calculation of the frequencies of attributes and attribute outcome pairs in the training data set.

IV PROBLEMS IN THE PREVIOUS RESEARCH

From the review of the past research it is noticed that many data mining techniques like SVM, CART, C4.5, ID3 algorithms, Least Square – Support Vector machine, Navie Bayies, FT Tree, K Start, Back propagation and hybrid methods were used in the existing model to diagnose hepatitis for better result and with minimum time period. Many filters and wrappers were used to compare the features. Existing model has given the results considering few parameters taken from the UCI repository only the models were not compared with the clinical data [4, 5].

Existing model has considered adult data [8, 9, 11] the diagnosis of the children is different form the adult, so the proposed research will consider the diagnosis of hepatitis for the children clinical data and the results will be compared with the UCI databases. In the existing model for the diagnosis of the hepatitis the missing parameters were totally omitted and the results were obtained. But the missing parameters consist of nearly about 50% of the data which are not considered. The prognosis algorithms of the existing models were comparing the results with very few classification algorithms only [5, 9]. The results were not compared with the other data mining techniques which are highly recommended. Majority of the cases the existing models used only few filters and wrappers [7] but it can also be compared with others. The experimental results were done only on WEKA [5, 6] but it can also be implemented on other available tools and the results can be compared.

V PROPOSED FRAMEWORK

The proposed framework is discussed below with the following steps. First step is to understand the application domain which is relevant to extract knowledge for achieving the final objective. Then the data needed to diagnosis hepatitis is to be collected and then the database to be prepared and mined for making appropriate decisions.

Data warehousing is an approach to store large volume of data for extracting pattern from that partitioning the data base into training and test dataset. Selecting the dataset, examine on subset of variables for determining the feasibility to solve the problem on which the discovery to be done. Clean the data for the training set for finding useful features to represent the data depending on the goal of the task.

Implementing dataset with different data mining algorithms then compare the data with different data mining techniques like is Neural Network, Bayesian, SVM, Logical regression. Select appropriate technique to act as a predicator. From the existing parameter and provide the additional and variation parameters to find out the best data set. Then validate each model by implementing with current cross-validation techniques which are used and providing the additional cross-validation techniques to find the model which gives maximum accuracy within a specified time. Finally select an optimal model to diagnose the hepatitis in children and predict the disease occurrence. Model framework for diagnosis of Hepatitis is given in structural diagrammatical presentation which is given below.

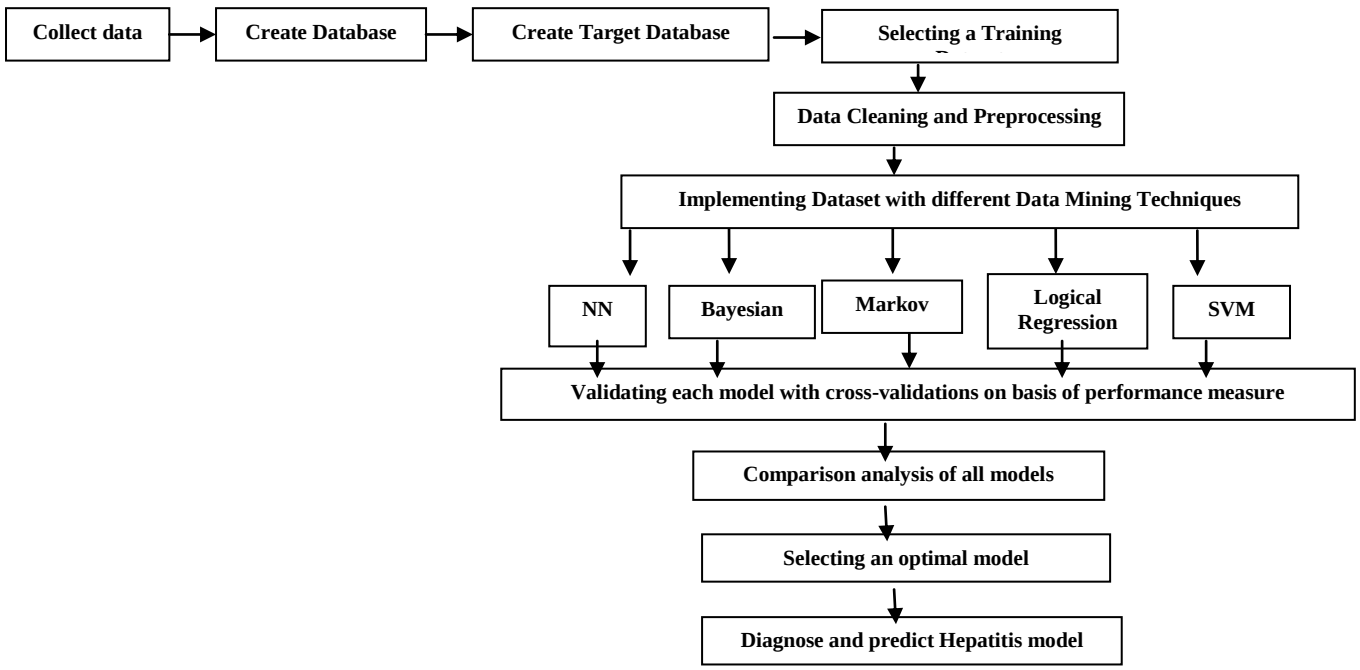

Fig. 1 Model framework for diagnosis of hepatitis

VI. CONCLUSION

Improving the performances for the diagnosis of hepatitis is the major task which can be performed by the data mining. This papers reviews different relevant data mining tools which are helpful for the hepatitis diagnosis. It specifies the accuracy rate implemented through that techniques. Here we have proposed that existing model can be still enhanced by considering above specified methods to achieve the better performances. The proposed model can also be used to predict the children hepatitis diagnosis.

Acknowledgement

The review papers in this research work were collected from the Vikram Sarabhai Library - Indian Institute of Management, Ahmedabad.

References

[1] W.M. Lee, *Hepatitis B virus infection*, N. Engl. J. Med. 337 (1997) 1733.
[2] J. Cohen, *The scientific challenge of hepatitis C*, Science 285 (1999) 26.
[3] Yılmaz Kaya, Murat Uyar, *A hybrid decision support system based on rough set and extreme learning machine for diagnosis of hepatitis disease.* 2013 Elsevier, 3429–3438
[4] Javad Salimi Sartakht, J. S. (2011). *Hepatitis disease diagnosis using a novel hybrid method*. Elsevier, 570-579.
[5] Duygu ÇCalisir, Esin Dogantekin, *A new intelligent hepatitis diagnosis system: PCA–LSSVM*, 2011 Elsevier, 10705–10708
[6] G.Sathyadevi, *Application of CART Algorithm in Hepatitis Disease Diagnosis*, 2011 IEEE, 1283-1287
[7] A.H.Roslina, & A.Noraziah. (2010). *Prediction of Hepatitis Prognosis Using Support Vector Machine and Wrapper Method*, IEEE, 2209-2211.
[8] Fadl Mutaher Ba-Alwi, H. M. (Volume 4, Issue 8, August-2013). *Comparative Study for Analysis the Prognostic in Hepatitis Data: Data Mining Approach*. International Journal of Scientific & Engineering Research, 680-685.
[9] S. B. Kotsiantis, *Increasing the Classification Accuracy of Simple Bayesian Classifier*, AIMSA, pp. 198-207, 2004
[10] Houzifa M. Hintaya, F. M.-A. (August-2013). *Comparative Study for Analysis the Prognostic in Hepatitis Data: Data Mining Approach*. International Journal of Scientific & Engineering Research, Volume 4, Issue 8, 680-685.
[11] Murat Uyar, Y. K. (2013). *A hybrid decision support system based on rough set and extreme*. Elsevier, 3429–3438.
[12] K. Polat, S. Gunes, *Hepatitis disease diagnosis using a new hybrid system based on feature selection (FS) and artificial immune recognition system with fuzzy resource allocation*, Digital Signal Processing 16 (6) (2006) 889–901.
[13] http://hepatitis.about.com/od/overview/a/numbers.htm

