IJOCIT
ISBN: 2345-3877
www.ijocit.org
Volume 4, Issue 3

Original Research

Presenting a Hybrid Model for Early Diagnosis of Hepatitis by Applying Data Mining Techniques

Peyman Bayat [1, *]
Masumeh Motevalli Almouti [1]

Received: 02 Jun 2016
Accepted: 19 Aug 2016

Abstract
Data mining is a science that works on mining of undiscovered patterns and relationships from raw data. These patterns are used for prediction or identification of events before their happening or at early stages of their happening. In recent decade, the application of data mining for quick diagnosis (in early stages of illness) of many diseases (such as: various kinds of cancers, diabetes, cardiovascular disorders and etc.) and also the prediction of the possibility of specific diseases by the discovering association relationships and the investigation of diseases-related factors. In this study, we tried to investigate hepatic diseases, their symptoms and their treatment. Then, some studies were reviewed that were related to data mining techniques for early detection of various diseases and different algorithms like: decision tree, SVM, Rough Set, neural network and Bayesian network that according to many studies done by various scientists around the world had the highest accuracy for detection of hepatitis. Later, a combinational approach was applied for the highest accuracy percentage of present approaches that in this method, a proposal combination of three algorithms: neural networks, Rough Set and Bayesian networks. The results of the proposal combinational method indicated that its efficiency and accuracy was higher in comparison to other methods.

Keywords: Data mining, Association Rules, Classification, Pattern Discovery, hepatitis Early Detection.

[1] Department of Computer Software Engineering, Rasht Branch, Islamic Azad University, Rasht, Iran
[*] Corresponding Author: drbayat.iau@gmail.com.

1. Introduction

In recent years synchronized with industrial development and modernity, the rate of some diseases like cancer, hepatitis, MS and etc. have been increased too. The physicians use various approaches like: blood and urine test, medical ultrasound, radiography, mammography and MRI to diagnose each of these diseases worldwide. In addition, in some cases, they are obliged to take some samples named as biopsy.

In spite of presence of these diagnostic methods, some of these diseases are not identified in early stages and they are mostly diagnosed when they are relatively developed, threatening the patient's life. So at the best possible state, it enforces high costs to the person and society for relieving or treatment. So, having some approaches are one of the essentials in medical science for pre-diagnosis according to some personal features (like heritage, blood group, weight, location, nutrition and etc.) or disease diagnosis in early stages and start in treatments for inhibition of disease development. In recent years, the use of data mining for diagnosis of the disease possibilities has been extended.

The most significant challenge for this subject is to choose the best technique of data mining for disease prediction. In this research, we will try to test different data mining techniques on the data of hepatic patients and at last we will introduce a hybrid model with the best technique and the highest level of accuracy for prediction of hepatitis.

2. Related Works

Data mining includes the application of some techniques for the discovery of information and new knowledge from extensive volume of pre-existing data. This information and knowledge were unknown previously and these were capable to be utilized for a work or a special aim. In recent decade, utilization of data mining techniques has been extended and it nearly has application in all platforms of human life (medicine, industry, agriculture, armed forces, space exploration and etc.).

In this part, we will investigate about some researches which have been conducted on data applications for pre-diagnosis, prediction or treatment of hepatitis.

Vijayarani & et al, have tried to use classification algorithms to predict hepatic diseases. In present study, SVM and Naïve Bayes have been applied. Their efficiency in terms of conducting time and classification accuracy were compared. The results revealed that SVM has higher accuracy for the prediction of hepatic diseases [1].

Hyontai Sung suggested a method based on over-sampling to form a decision tree for the diagnosis of hepatic diseases with high accuracy in minimum classes to recover effective data. The finding about the algorithms of C4.5 and CART decision trees confirmed the accuracy and validity of the method [2].

Shubpreet Kaur & et al [3] have reviewed various techniques of data mining for diagnosis and predictions in medicine and finally they worked on future tendencies which are going to be focused on data mining techniques and knowledge discovery in order to help people in medical issues by means of data mining tools. Furthermore, they worked on total challenges related to medical issues that it can find some answers to them.

T. Karthikeyan and et al [4] have studied data related to hepatic patients from UC Irvine database. They used different classification models including: Bayes Net, Random forest Bayes. Naive Bayes, J48, Naïve Bayes Updatable, Multi-Layer Perceptron and the results were investigated in terms of conducting time and accuracy. At last, they claimed that the efficiency of Naïve Bayes algorithm for the diagnosis of hepatic patients was better than other classification models.

Pei Shen and et al [5] had applied three general data mining techniques (including: decision tree, SVM artificial neural networks and machines) for analysis of risk factors in treatment of A hepatic patients. Their results analysis showed that SVM has the highest accuracy rate in prediction of the disease.

Fadl Mutaher and et al [6] investigated different classification algorithms including Naïve Bayes Updatable, Naïve Bayes Neural network, LMT, J48, KStar, FT tree for data analysis of hepatitis. The classification results in terms of conducting time and accuracy of algorithms have revealed that the efficiency of Naïve Bayes classification was better that other techniques launched on hepatic patients data base in terms of accurate diagnosis or treatment.

Kwong-Sak Leung and et al [7] presented a data mining frame which includes evaluation of molecular analysis, clustering, feature selection and classification techniques. They gathered HBV DNA samples and C and B genes from 200 patients. The potential markers were selected according to achieved information from other classifications during feature selection. A classification method was suggested by integer non-linear. The best efficiency was applied for this proposed method by fuzzy logic and integer non-linear method. The results of evaluation of these classification methods revealed that their accuracy and sensitivity are 70% and 80% for liver cancer diagnosis respectively.

Huda Yasin and et al [8] searched on the responsible factors for C hepatitis. California University which included 9 features and 1 property named as class. In addition, these data included 155 records. Normalizing techniques were applied for repair and insertion of faulty or blank fields.

Binary Logistic Regression was utilized for classification of patients by quantitative and qualitative methods in order to minimize the size of data collection. The results showed that the accuracy of classification was 89%. In this proposed method, the complexity of these features was less and only by 37% of all data, this method can create a pattern.

3. The Proposed Hybrid Method

The proposed general framework in this research is depicted in figure 1 for diagnosis of hepatic patients in early stages.

The collected data are from UCI database in this study included the records of 200 patients with B hepatitis. There are 20 features for each patient in these data that the name of these features and the kind of acceptable amounts are mentioned in table 1. As it is seen in figure, all data of this thesis are classified into one of two kinds of numeric and categorical variants.

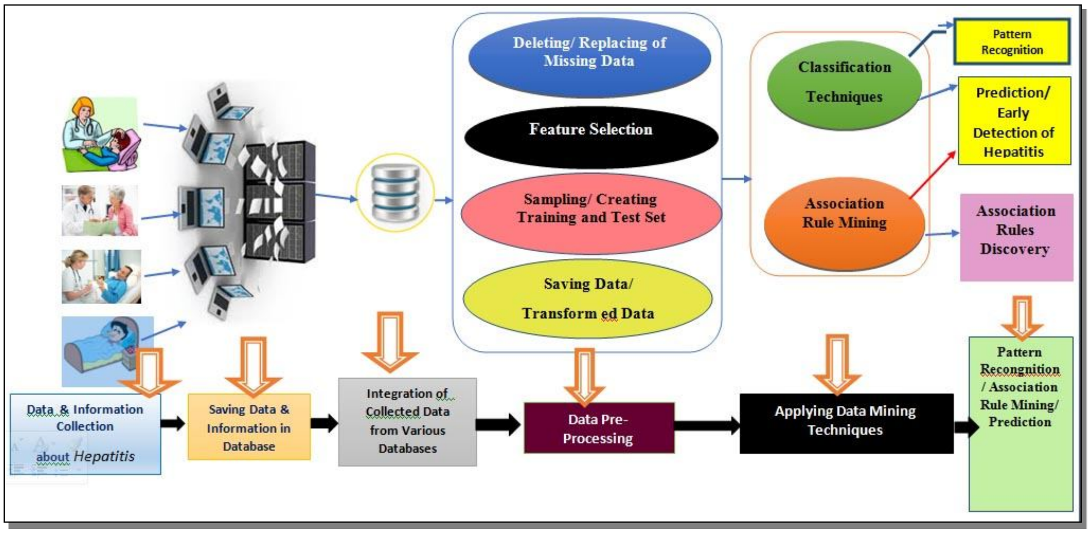
Deleting/ Replacing of Missing Data
Feature Selection
Sampling/ Creating Training and Test Set
Saving Data/ Transform ed Data
Classification Techniques
Association Rule Mining
Pattern Recognition
Prediction/ Early Detection of Hepatitis
Association Rules Discovery
Data & Information Collection about Hepatitis
Saving Data & Information in Database
Integration of Collected Data from Various Databases
Data Pre-Processing
Applying Data Mining Techniques
Pattern Recognition / Association Rule Mining/ Prediction

Figure 1. Our Proposed Framework for Early Detection/Prediction of Hepatitis

Table 1. Features in the Collected Data

Attribute	Value	Value Type	Attribute	Value	Value Type
Class	Die (1), Live (2)	Categorical	Malaise	No (1), Yes(2)	Categorical
Age	10, 20, 30, etc.	Numeric	Anorexia	No (1), Yes (2)	Categorical
Gender	Male (1) , Female (2)	Categorical	Liver Big	No (1), Yes(2)	Categorical
Steroid	No (1), Yes(2)	Categorical	Liver Firm	No (1), Yes (2)	Categorical
Antiviral	No (1), Yes (2)	Categorical	Spleen Palpable	No (1), Yes(2)	Categorical
Fatigue	No (1), Yes(2)	Categorical	Spiders	No (1), Yes (2)	Categorical
Bilirubin	0.39, 0.80, 1.20, 2.0, 3.0, 4.0, continues values	Numeric	Ascites	No (1), Yes(2)	Categorical
Alk Phosphate	33, 80, 120, 160, 200, 250	Numeric	Varices	No (1), Yes (2)	Categorical
Serum Glutamic-Oxaloacetic Transaminase	13, 100, 200, 300, 400, 500	Numeric	Albumin	2.1, 3.0, 3.8, 4.5, 5.0, 6.0	Numeric
Protime	10, 20, 30, …, 90	Numeric	Changes in Texture	No, Yes	Categorical

The stages of data analysis are:

- **Data Pre-Processed Stage:** For data analysis, it is vital to eliminate blank fields or filled with suitable amounts by present techniques.

By performing described methods in fourth chapter, we recover and fill/delete the blank fields. Table 2 shows the percentage of blank/missing fields for each one of features in collections of collected data.

Table 2. The Percentage of Missing Fields for Each One of Features

Attribute	% of Missing Data	Attribute	% of Missing Data
Age	0%	Age	0%
Antiviral	0%	Steroid	1%
Malaise	1%	Fatigue	1%
Liver Big	6%	Anorexia	1%
Spleen Palpable	3%	Liver Firm	7%
Ascites	3%	Varices	3%
Spiders	4%	Bilirubin	3%
SGOT	3%	Alk Phosphate	12%
Changes in Texture	0%	Albumin	27%
		Protime	43%

In addition, data normalization, selection of suitable data for analysis out of present features and finally the transformation of data in necessary cases are in pre-processes stage.

- **Early Data Analysis Stage:** In this stage, we examine whether there is a relationship between different features and hepatitis or not?
  - **The relationship between age and the rate of hepatitis level:** The most death cases belong to the patients whom age between 40- 50. While, the smallest amount of patients and death rate are under 20 years old.
  - **The relationship between anorexia and sex of hepatic patients:** In all age ranges, the majority of patients have symptoms of anorexia.
  - **The relationship between hepatomegaly and age:** Ignoring the patient's age, most of them have hepatomegaly symptoms.
  - **The relationship between bilirubin and death rate of the patients:** The increased rate of death cases is related to those people whom their bilirubin is higher than 2.5.
  - **The relationship between death rate and SGOT:** The highest rate of death is for patients with SGOT more than 50 and less than or equal with 100.
  - **The relationship between sex and patient death rate:** Sex is a significant factor in death determination because women have the lowest rate of death while most of the death cases are for male patients.

- **Pattern Recognition Stage:** We used different techniques to create patterns such as:
  - Decision Tree
  - SVM
  - Neural Network
  - Naïve Bayes
  - Rough Set
  - Regression

In addition, we used those algorithms related to association rules discovery.

3.1. Our Proposed Method

We will combine those algorithms that shoed the highest accuracy in our analysis with each other. This selective combinational method is summarized as fellow chart of Figure 2.

According to suggested method if all of three algorithms showed the person as a healthy or unhealthy person, 100% the person is healthy but if it is not like this , we will accept the answer if at least one of two suggested algorithms indicate person as healthy or unhealthy.

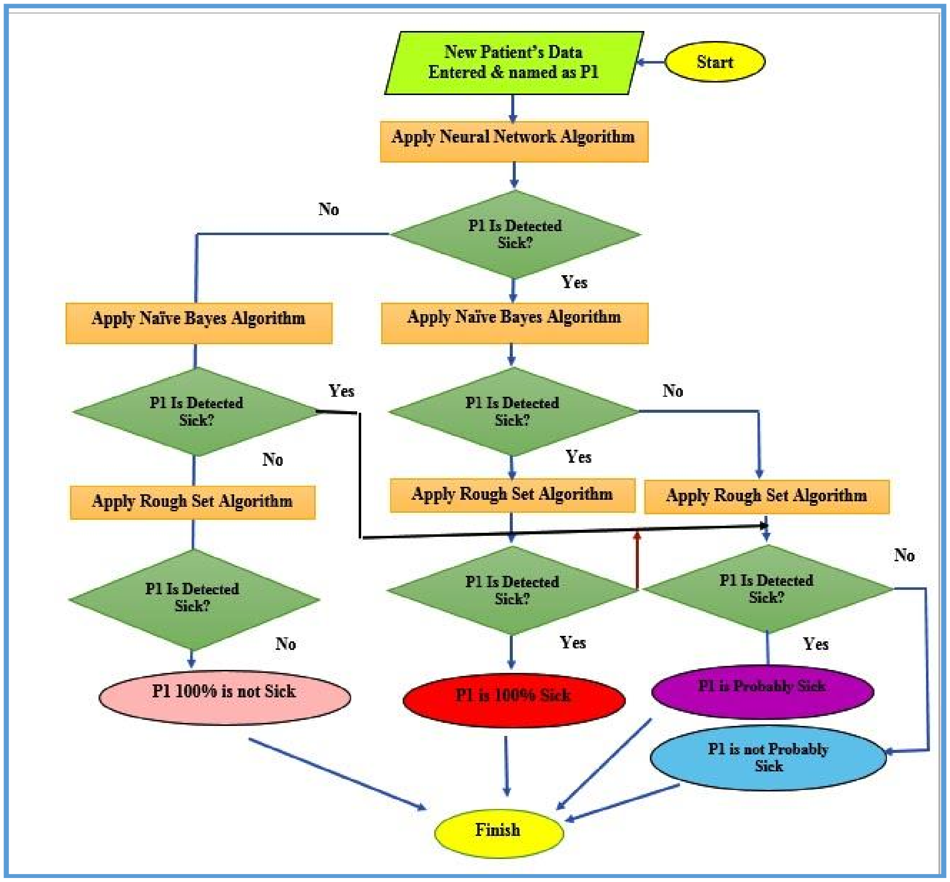

Figure 2. The suggested algorithm for Early Detection of Hepatitis

4. Evaluation of the Efficiency and the Accuracy of the Suggested Method

We will use the data of tests sets to compare the accuracy of algorithms in section 3.1. To do this, we will use the following formula:

- **Accuracy:** The meaning of accuracy calculation is the measurement of results quality in comparison with reality. The following formula is used for calculation of the applied algorithms accuracy:

$$Ac = \left(\frac{N_{TP} + N_{TN}}{N_{TP} + N_{TN} + N_{FP} + N_{FN}}\right)$$

(Accuracy Calculation)

In this formula:

TP is the number of the samples which have been diagnosed positive correctly. In other word, it is the number of patients and diagnosed as patient.

TN is the number of the samples which have been diagnosed negative correctly. In other word, it is the number of healthy individuals and who have been diagnosed as healthy persons.

FP is the number of samples which have been diagnosed positive incorrectly. In other word, those who are healthy but are diagnosed patient.

FN is the number of samples which have been diagnosed negative incorrectly. In other word, those who are healthy but are diagnosed healthy.

- **Accuracy + Sensitivity (F-Measure):** For this the following formula can be used:

$$FMeasure = \frac{2N_{TP}}{2N_{TP} + N_{FP} + N_{FN}}$$

- **Kappa:** Shows accuracy of achievement to the desired expectations.

$$Kappa = \frac{2(N_{TP}N_{TN} + N_{FN}N_{FP})}{(N_{TP}+N_{FN})(N_{TN}+N_{FN})+(N_{TN}+N_{FP})(N_{TP}+N_{FP}}$$

- **Precision:** Its goal is the measurement of performance quality that the results are achieved by it.

$$\text{precision} = \frac{\text{number of true positives}}{\text{number of true positives} + \text{false positives}}$$

The results are mentioned in Table 3.

Table 3. The Comparison of Evaluation Results of Conducted Methods for Pattern Recognition

Algorithm	Accuracy	Sensitivity	F-Measure	Kappa	Precision
Decision Tree	75%	79.9%	78.8%	73.3%	81.3%
SVM Algorithm	77.%	83.4%	81.3%	75.5%	83.1%
SVM Algorithm (Poly Function)	77.%	83.4%	80.3%	75.5%	83.1%
Neural Network Algorithm	88.%	85.7%	87.1%	86.1%	87.4%
Naïve Bayes	87.%	83.9%	85.6%	84.6%	84.1%
Rough Set	89.%	83.6%	86.7%	85.1%	89.9%

As it is seen in Table 3, three algorithms: neural networks, Bayesian and Rough set have the highest rate of accuracy among other algorithms. So our suggested method is presented for accuracy enhancement in the prediction of up-coming results. The results of conducted proposal method on 50 cases of test set is revealed in Table 4.

Table4. The Evaluation Result of Proposed Combinational Method

Algorithm	Accuracy	Sensitivity	F-Measure	Kappa	Precision
Proposed Hybrid Method	90%	86.3%	88.8%	87.7%	90.1%

As it can be understood from the comparison of Table 3 and Table 4, all results are optimized in our suggested approach. So this method can be used practically as an approach for optimization of diagnosis rate or hepatitis prediction.

5. Conclusion

Desired data are gathered for analysis from UCI hepatic diseases research center. Various algorithms like: decision tree, SVM, Rough Set, neural networks and Bayesian that according to many investigations done by different researchers worldwide that have the highest rate of accuracy in hepatic diagnosis were conducted on collected data. The results and achieved pattern finally was tested on the data of test sets.

At the next stage, a combinational method was presented for the highest percentage of accuracy of present methods that in this method, three algorithms: neural networks, Rough Set and Bayesian networks. The results of the evaluation of proposed combinational method showed that the efficiency and accuracy of the proposed method was higher than other methods. So the proposed method was applied for diagnosis or prediction of hepatitis as a help for physicians.

6. References

[1] S. Vijayarani, S. Dhayanand, (2015), “Liver Disease Prediction using SVM and Naïve Bayes Algorithms”, International Journal of Science, Engineering and Technology Research (IJSETR), Vol. 4, No. 4, pp. 817-820.

[2] Hyontai Sug, (2012), “Improving the Prediction Accuracy of Liver Disorder Disease with Oversampling”, Applied Mathematics in Electrical and Computer Engineering, pp. 331-335.

[3] Shubpreet Kaur and Dr. R.K.Bawa, (2015), “Future Trends of Data Mining in Predicting the Various Diseases in Medical Healthcare System”, International Journal of Energy, Information and Communications, Vol.6, No.4, pp.17-34.

[4] T.Karthikeyan & P.Thangaraju, (2013), “Analysis of Classification Algorithms Applied to Hepatitis Patients”, International Journal of Computer Applications, Vol. 62, No.15, pp. 25-30.

[5] Pei Shen and Jikai Zhang, “International Journal of Hybrid Information Technology, (2015), Vol.8, No.4, pp. 193-200.

[6] Fadl Mutaher Ba-Alwi, Houzifa M. Hintaya, (2013), “Comparative Study for Analysis the Prognostic in Hepatitis Data: Data Mining Approach”, International Journal of Scientific & Engineering Research, Volume 4, No. 8, pp. 680-685.

[7] Kwong-Sak Leung, Kin Hong Lee, Jin-Feng Wang, Eddie Y.T. Ng, Henry L.Y. Chan, Stephen K.W. Tsui, Tony S.K. Mok, Pete Chi-Hang Tse, and Joseph Jao-Yiu Sung, (2011), “Data Mining on DNA Sequences of Hepatitis B Virus”, IEEE/ACM Transactions On Computational Biology And Bioinformatics, VOL. 8, NO. 2, pp. 428-440.

[8] Huda Yasin, Tahseen A. Jilani & Madiha Danish, (2011), “Hepatitis-C Classification using Data Mining Techniques”, International Journal of Computer Applications, Vol. 24, No.3, pp. 1-6.