International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-5 Issue-6, August 2016

Data Mining using Meta Heuristic Approaches for Detecting Hepatitis

Neenu R S, Greeshma G Vijayan

Abstract*—*Clinical Data Mining involves the process of extracting, analyzing and finding the available data for clinical decision making. Mining data from clinical data set is not an easy task as they are inserted manually. In this paper, a solution for accurately predicting the presence or absence of hepatitis is proposed. The proposed technique is applied on clinical data sets taken from University of California at Irvine (UCI) machine learning repository. The proposed system contains two main subsystems for preprocessing and classifying. In the preprocessing subsystem the missing values in the data set is handled using missing data imputation methods like litwise deletion or mean/mode imputation method. If the percentage of missing values in a tuple is greater than 25%, then the tuple is rejected from the dataset else it was imputed by the most frequently used value. After handling the missing value, the relevant attributes are selected using meta-heuristic approaches like Particle Swarm Optimization (PSO) is used for feature selection. The reducts obtained after preprocessing are fed into the classification. In the classification subsystem the selected reducts are trained and tested using back propagation neural network. This paper aims at accurate prediction of diseases by analyzing clinical data sets.

Index Terms*— *Back propagation neural network, Clinical Data Mining, Particle Swarm Optimization (PSO), University of California at Irvine (UCI).

I. INTRODUCTION

Data mining is the process of discovering useful knowledge or patterns from large datasets. Nowadays there is huge amount of data being collected and stored in databases everywhere across the globe. There is invaluable information and knowledge hidden in such databases; and without automatic methods for extracting this information it is practically impossible to mine them [1]. Data mining have a very important place in medical science in order to enhance the decision making and diagnosis process. The amounts of clinical datasets, which contain details patients, are increasing day by day. These health-care data is maintained in repositories but the problem is on mining knowledge from such clinical data. The extracted knowledge should be novel, interesting, precise and comprehensible so as to improve the decision making process. Efficient and robust computational algorithms are required to develop an optimized decision making model. The widespread availability of new computational methods and tools for data analysis and predictive modeling requires medical researchers and practitioners to systematically select the most appropriate strategy to deal with clinical prediction problems. Clinical Data Mining (CDM) is an exact solution for mining knowledge from clinical data repositories. CDM involves the conceptualization, extraction, analysis and interpretation of available clinical data for practice knowledge building, clinical decision making and practitioner reflection [2].

In this work a classifier is developed which will guess predicts whether the patient is infected with hepatitis or not. Particle Swarm Optimization (PSO) and back propagation learning algorithm is used for developing the classifier. Particle swarm optimization (PSO), proposed by Dr. Eberhart and Dr. Kennedy in 1995, is a population based stochastic optimization technique. It is a population-based search algorithm and is initialized with a population of random solutions, called particles [3]. In this work PSO is used for selecting attributes from the clinical data set and these selected attributes are given as input for the back propagation neural network. Back propagation (BP) is commonly used for training artificial neural network (ANN) model proposed by Rumelharth. Back propagation neural network (BPNN) consists of at least three types of layers of units, input layer, at least one intermediate hidden layer, and output layer. In BPNN units are connected in feed-forward fashion with input units fully connected to hidden layer units which are fully connected to units in output layer. The output of a BPNN is interpreted as a classification decision [4].

The proposed technique has been applied on hepatitis datasets obtained from University of California at Irvine (UCI) machine learning repository. Hepatitis refers to inflammation of liver, caused by viral infection. The hepatitis virus can be grouped as A, B, C, D, and E. These viruses cause acute to chronic disease; type B and type C, especially, lead to chronic diseases.

II. RELATED WORKS

Kaya and Uyar [5] has proposed a hybrid approach for the diagnosis of hepatitis based on rough set (RS) and extreme learning machine (ELM). This system mainly consists of two stages. By using the RS method redundant features are removed from the data set in the first stage. In the second stage classification has been done using ELM. In order to test the model hepatitis data set taken from UCI repository is used. In this method the missing values are removed after selecting the reducts because data set includes more missing values. After handling missing values classification process has performed through EML. They produced 20 reducts using RS.

Manuscript published on 30 August 2016.
***** Correspondence Author (s)
Neenu R S, M.Tech Scholar ,Department of Computer Science and Engineering, LBS Institute of Technology for Women, Thiruvananthapuram, India.
Greeshma G Vijayan, Assistant Professor, Department of Computer Science and Engineering, LBS institute of Technology for Women, Thiruvananthapuram, India.

The researches have obtained 100% accuracy when training and testing data sets were selected at the rate of 80-20%. By using Support vector mechanism and simulate annealing (SVM-SA), Saratakhti et al. [6] developed a hybrid system for detecting hepatitis disease. SVM are supervised learning models which will analyze data used for classification and regression analysis [7]. It is also a well-known algorithm for disease diagnosis. Simulated annealing (SA) is a probabilistic technique for approximating the global optimum of a given function. The hepatitis data sets were taken from UCI. The class label Die has 32 cases and live has 123 cases. The proposed method includes following stages: Data Preprocessing, Scaling and Simulated Annealing. In order to enhance the classification accuracy tuned parameters were used. The accuracy obtained from this model is 96.25%.

Çalişir and Dogantekin [8] proposed an intelligent system using principal component analysis (PCA) and least square support vector machine (LSSVM). PCA, invented by Karl Pearson in 1901, is a technique used to emphasize variation and bring out strong patterns in a dataset. It's often used to make data easy to explore and visualize. PCA is sensitive to the relative scaling of the original variables. LSSVM are a set of related supervised learning methods that analyze data and recognize patterns, and which are used for classification and regression analysis. The proposed system mainly consists of two stages. In the first stage feature extraction and feature reduction was done using PCA and in the second stage classification was done by using LSSVM classifier. The hepatitis disease data sets are obtained from UCI repository. Initially there were 19 feature attributes and after feature reduction and extraction using PCA it was reduced to 10 attributes. These 10 attributes are given as input to the classifier. The researches have obtained an accuracy of 96.12%.

Nahato et al. [9] have built a classifier that will predict the presence or absence of a disease by learning from the minimal set of attributes that has been extracted from the clinical dataset. The researchers used rough set indiscernibility relation method with back propagation neural network (RS-BPNN). The missing values in the data set are handled first to obtain a smooth data set and selection of appropriate attributes from the clinical dataset by indiscernibility relation method. Then the selected reducts are classified is using back propagation neural network. The classifier has been tested with hepatitis, Wisconsin breast cancer, and Statlog heart disease datasets obtained from the University of California at Irvine (UCI) machine learning repository. The accuracy obtained from the proposed method is 97.3%, 98.6%, and 90.4% for hepatitis, breast cancer, and heart disease, respectively. The proposed system provides an effective classification model for clinical datasets.

III. PROPOSED SYSTEM

The proposed system mainly consists of two subsystems. First one is called preprocessing subsystem and the second one is the classification subsystem. Fig. 1 shows the architecture of proposed technique.

A. Dataset Description

For this work clinical datasets of hepatitis have been selected from the UCI [10] machine learning repository. Hepatitis dataset consists of 155 samples with 19 case attributes and a class label. Class labels in the datasets are to predict whether the patient with this disease will live or die. The dataset has 32 Die and 123 Live instances. Table I describes the attributes of the hepatitis dataset.

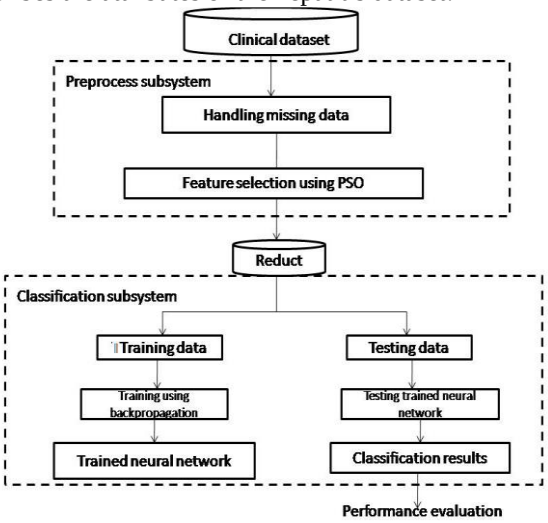

Fig.1: Proposed System Architecture

B. Handling Missing Values

Since the clinical datasets are inserted manually the chance of having missing values are high. In order to get an accurate result and to make the preprocessing step easier we have to avoid these missing values. Lot of methods is available for handling missing values like litwise deletion, mean/mode imputation and k-nearest neighbor. Litwise method is the simplest way of handling missing data in which the attribute containing missing values are deleted. In the Mean/Mode imputation method the missing values for a given attribute are replaced by the mean of all known values of that attribute in the class where the missing attributes belongs.

Table I. Description of hepatitis dataset

Number	Attribute name	Domain Values	Number of missing values
1	Age	10, 20, 30, 40, 50, 60, 70, 80	0
2	Sex	Male, female	0
3	Steroid	No, yes	1
4	Antivirals	No, yes	0
5	Fatigue	No, yes	1
6	Malaise	No, yes	1
7	Anorexia	No, yes	1
8	Liver big	No, yes	10
9	Liver firm	No, yes	11
10	Spleen palpable	No, yes	5
11	Spiders	No, yes	5
12	Ascites	No, yes	5
13	Varices	No, yes	5
14	Bilirubin	0.39, 0.80, 1.20, 2.00, 3.00, 4.00	6
15	Alk phosphate	33, 80, 120, 160, 200, 250	29
16	Sgot	13, 100, 200, 300, 400, 500	4
17	Albumin	2.1, 3.0, 3.8, 4.5, 5.0, 6.0	16
18	Protime	10, 20, 30, 40, 50, 60, 70, 80, 90	67
19	Histology	No, yes	0
20	Class	Die, live	0

In this proposed system the missing values are handled as follows. If the percentage of missing value in a tuple is greater than or equal to 25% then we use litwise method i.e. we will reject that tuple from the dataset. Otherwise we will go with the mean/mode imputation method where the missing values are imputed by the most frequent values of the attribute in the class that belongs to the tuple.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

After handling the missing values the hepatitis dataset is reduced to 147 samples with 18 attributes. From the dataset the Protime attribute and 8 tuple were deleted.

C. Feature Selection using PSO

Feature selection is used to choose a subset of input variables by eliminating features with little or no predictive information using meta-heuristic approach [11]. The main objective of feature selection is to select a minimal subset of features according to some reasonable criteria.

PSO is a population-based stochastic optimization technique, which simulates the social behavior of organisms, such as bird flocking and fish schooling, to describe an automatically evolving system. In PSO the potential candidates are known as particles [12]. In PSO, possible particles move in the search space to find the global optimum. This movement is based on proper combinations of control parameters and a replacement formula. Particles also change their positions using their own position called P_{best} (particle best) and swarms best position G_{best} (global best). When a better solution is discovered by any particle, all particles improve their positions to this better solution in the search space. Fig. 2 shows the flow chart of PSO

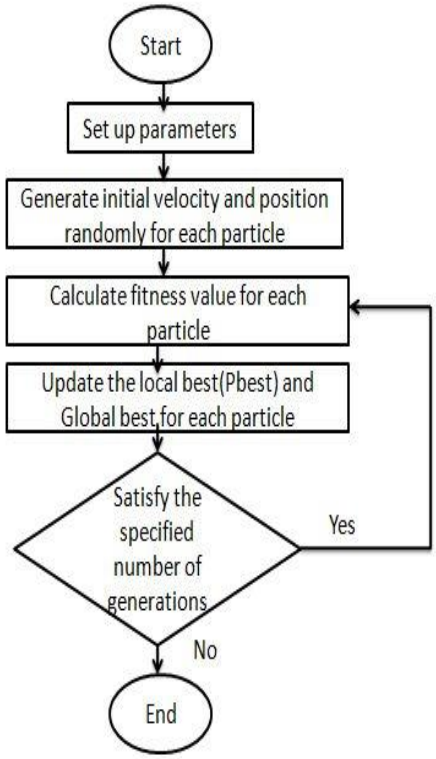

Fig 2. Flow chart of PSO

The main steps of PSO algorithm is as follows.

Step 1. Initialize the particles X_i with initial random positions in search space, and the velocities of particles V_i in a given range randomly and define these as the best known positions (P_{best}) of each particle.

Step 2. Define the objective function f that needs to be optimized.

Step 3. For each particle calculate the distance to the optimal solution called fitness, and apply equation (1):

$$\text{If} \quad f(X_i) < f(P_{best}) \; P_{best} = X_i \qquad (1)$$

Step 4. Select $global_{best}$ among the P_{best}

Step 5. If $f(global_{best})$ reaches to the optimal solution, terminate the algorithm. Otherwise; update the velocity V_i and the particles X_i according to given equations:

$$V_i = \omega \times V_i + c_1 .(P_{best} - X_i) + c_2 .(global_{best} - X_i) \qquad (2)$$

$$X_i = X_i + V_i \qquad (3)$$

Step 6. Go to *Step 3*, if stopping criteria are not satisfied.

Where c_1 is called cognitive parameter and c_2 is the social parameter. The selected subset of attributes is referred as reduct (Red(R)) [9]. The minimal subset of attributes with the same property as that of whole attribute is called reduct (Red (R)). The intersection of the elements of reducts is called core (C). By excluding empty set and whole conditional attribute, the total number of minimal subsets of attributes (S) competing for reducts becomes

$$S = 2^n - 2 \qquad (4)$$

Where n stands for total number of attributes.

D. Training and Testing the Classifier

The output after feature selection is given as the input for the BPNN for testing and training the classifier. Fig 3. shows the architecture of BPNN.

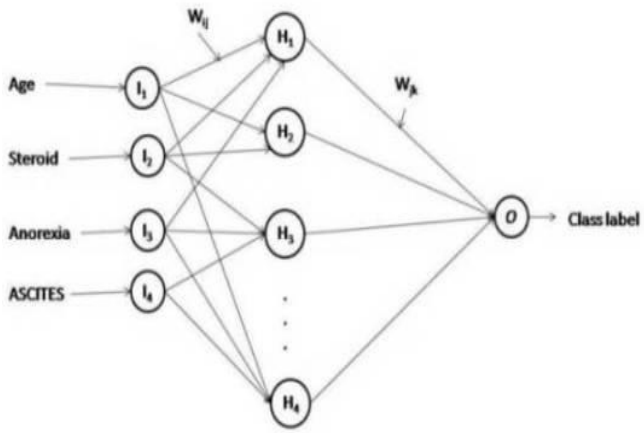

Fig 3. Architecture of BPNN

After feature selection we will get a confusion matrix as output. The confusion matrix will consist of true positive (TP), false positive (FP), true negative (TN), and false negative (FN) values. The accuracy of the back propagation neural network is computed using the confusion matrix. The percentage of sample data that are correctly classified by the classifier is called as accuracy. Accuracy can be calculated as follows

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100$$

IV. Experimental Evaluation

Eclipse IDE (4.3.0 release) and Java is used to implement the proposed system. The data set used is Hepatitis data set from UCI repository [11]. The data set contains 155 samples with 19 case attributes and a class attribute. The class attribute will predict whether the patient with hepatitis will live or die. PSO is used to extracting the feature attributes from the data set after handling the missing values.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

These reducts are linked with the input layer of BPNN. The BPNN is used for testing and training the system.

Hepatitis dataset has a set of 262142 attributes which are competing for reducts. The selected reducts are shown in table II.

Table II. Selected reducts of hepatitis dataset

Reduct	Number of attributes	Attribute set
R1	9	[1 0 1 0 1 0 0 1 1 0 0 0 0 1 1 1 1 0]
R2	10	[1 1 1 0 1 0 0 1 1 0 0 0 0 1 1 1 1 0]
R3	10	[1 0 1 0 1 0 0 1 0 1 1 0 0 1 1 1 1 0]
R4	11	[1 0 1 0 1 1 0 1 1 0 1 0 0 1 1 1 1 0]
R5	11	[1 0 1 0 1 1 0 1 1 1 0 0 0 1 1 1 1 0]
R6	12	[1 0 1 0 1 0 1 1 0 0 1 1 1 1 1 1 1 0]
R7	12	[1 0 1 0 1 0 1 1 0 1 1 0 1 1 1 1 1 0]
R8	13	[1 1 1 1 1 0 0 0 0 1 1 1 0 1 1 1 1 1]
R9	13	[1 0 1 1 1 0 0 1 1 0 1 0 1 1 1 1 1 1]
All	18	[1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1]

The proposed system gives best result for the data division of 80-20 where all the selected reducts obtain accuracy of more than 90%. The ROC graph of LIVE class value is shown in Fig 4. The X-axis shows the false positive rate and the Y-axis shows the true positive rate. Fig 5. Shows the ROC curve for class value DIE.

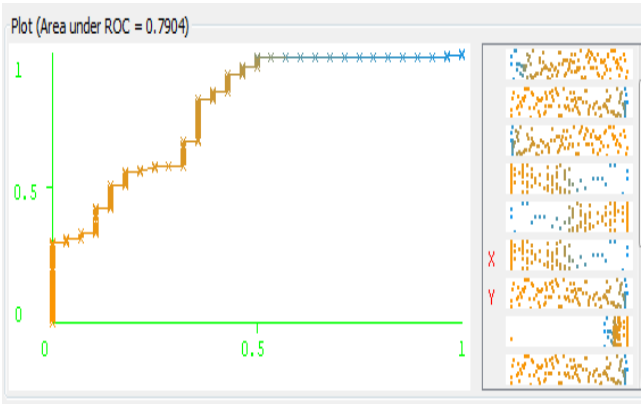

Fig 4. ROC curve for LIVE class.

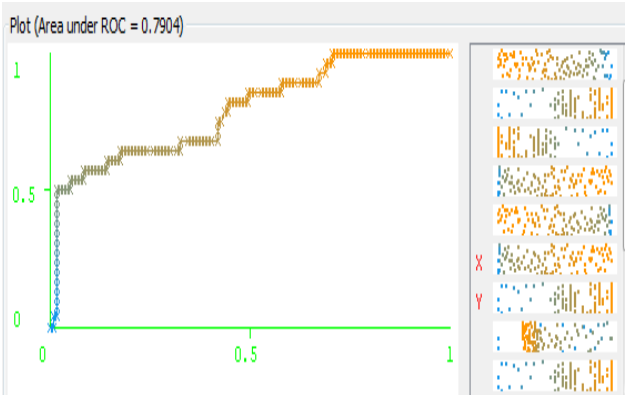

Fig 5. ROC curve for DIE class

Fig 6. Shows the performance of target class and output result by taking first 100 inputs.

V. CONCLUSION

The role of clinical data mining in the field of medical science is increasing day by day. More automatic methods are required for enhancing the decision making and diagnosis of disease easier. This paper combines particle swarm optimization with back propagation neural network to analyze the clinical dataset. The proposed method is applied in hepatitis dataset from UCI. PSO is used for feature selection. Before applying PSO the missing values in the datasets are handled either by litwise deletion or mean/mode imputation method. The output after feature selection is linked with the input layer of BPNN. Backpropagation neural network is used for classification and to predict whether the patient with hepatitis will live or die. The experimental results demonstrate that the proposed techniques are successful. The accuracy obtained from the proposed method is 94.92. It is possible to apply the same method for diagnosing other diseases by giving their datasets instead of hepatitis dataset. In the future, these different meta-heuristic approaches like Ant Colony Optimization, Honey Bee Optimization, Lion Optimization etc can be used in the place of PSO.

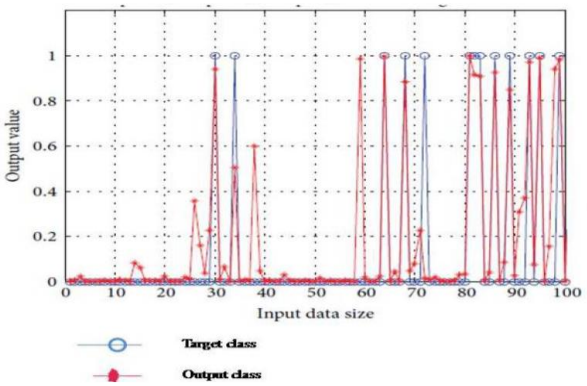

Fig 6. Hepatitis sample dataset performance

ACKNOWLEDGMENT

We are greatly indebted to our principal, Dr. JAYAMOHAN J, Dr. V. GOPAKUMAR, Professor, Head of the Department of Computer Science and Engineering, Mrs. GREESHMA G. VIJAYAN, Assistant Professor, Department of Computer Science and Engineering, LBS Institute of Technology for Women who have been instrumental in keeping my confidence level high and for being supportive in the successful completion of this paper. We would also extend our gratefulness to all the staff members in the Department; also thank all my friends and well-wishers who greatly helped me in my endeavor. Above all, we thank the Almighty God for the support, guidance and blessings bestowed on us, which made it a success.

REFERENCES

1. Fabricio Voznika and Leonardo Viana, "Data Mining Classifications".
2. What is clinical dataming? http://www.slideshare.net/empowerbpo/what-is-clinical-data-mining
3. Yamille del Valle, Ganesh Kumar Venayagamoorthy, Salman Mohagheghi, Jean-Carlos Hernandez, and Ronald G. Harley "Particle Swarm Optimization: Basic Concepts, Variants and Applications in Power Systems", IEEE Transactions On Evolutionary Computation, VOL. 12, NO. 2, APRIL 2008
4. R. C.Chakraborty, "Back Propagation Network: Soft Computing Course Lecture", 15-20, Aug 10,2010.
5. Y. Kaya and M. Uyar, "A hybrid decision support system based on rough set and extreme learning machine for diagnosis of hepatitis disease", ApplieSoft Computing Journal, vol. 13, no. 8, pp. 34293438, 2013.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

6. J. S. Sartakhti, M. H. Zangooei, and 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, vol. 108, no. 2, pp. 570579, 2012.
7. Support Vector Mechanism.- https://en.wikipedia.org/wiki/Support_vector_machine
8. D. Çalişir and E. Dogantekin, “A new intelligent hepatitis diagnosis system: PCALSSVM”, Expert Systems with Applications, “vol. 38, no. 8, pp. 1070510708, 2011.
9. Kindie Biredagn Nahato, Khanna Nehemiah Harichandran and Kannan Arputharaj, “Knowledge Mining from Clinical Datasets Using Rough Sets and Backpropagation Neural Network”, Hindawi, 2015
10. K. Bache and M. Lichman, UCI Machine Learning Repository, University of California, School of Information and Computer Science, Irvine, Calif, USA, 2013.
11. Hany M. Harb, and Abeer S. Desuky , “ Feature Selection on Classification of Medical Datasets based on Particle Swarm Optimization “, International Journal of Computer Applications (0975 – 8887) Volume104– No.5, October 2014.
12. Ezgi Deniz Ülker and Sadık Ülker, “Application of Particle Swarm Optimization To Microwave Tapered Microstrip Lines”, Computer Science & Engineering: An International Journal (CSEIJ), Vol. 4, No. 1, February 2014.