ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 4, Issue 1, January 2016

Eamining and Comparing Data Mining-Based Techniques for Hepatitis Diagnosis

Mohammad reza moshkani[1], Mahdi Rousta[2]*, Dr.Yaghoub Farjami[3]

Associate Professor, Department of Biology, Qom University, Qom, Iran[1]

M.S. Student, Department of Information Technology Engineering, Qom University, Qom, Iran [2]

ABSTRACT: Increasing advances in information technology has led to significant growth in sciences. One of the fields in which significant changes has occurred is the medical field. Using data- mining techniques in this branch of science has helped physicians in all subjects, in particular diagnosis of sicknesses. Hepatitis diagnosis is highly difficult due to limited clinical diagnosis of the disease in its early stages. To this end, this paper tries to introduce and recommend the best way to diagnose hepatitis as well as to compare common clustering methods such as decision trees, neural networks, and SVM. Evaluation criteria of classification methods are the accuracy of each of methods and Clementine software along with data base in the University of California has been used to test each method. Obtained results show that neural network algorithm enjoys higher accuracy in comparison with other algorithms. Using neural network algorithm can accurately predict 89.74% hepatitis.

KEYWORDS: Hepatitis diagnosis; data mining; decision tree; neural network; SVM.

I. INTRODUCTION

The expansion of computer usages in business activities has led to the rapid growth of databases and data gathering by most of organizations. Every day, a huge amount of generated data is stored in different data centers. In recent years, the tendency to search for detection of replicated patterns has considerably increased in order to improve decision-making. Medical and health fields are among important measures in industrial societies. Knowledge extraction among massive related data to history of sickness and medical files by the help of data-mining process can lead to identification of rules governing the creation, growth, and illness spread, providing valuable information for specialists and health officials to identify the reasons of occurrence, prediction, and treatment considering environmental factors. The result of this issue means increased life span and peace for society members. [1].

Based on the type of knowledge being extracted, data-mining technique is divided in to three categories including pattern classification, data clustering, and extraction of associative rules [3]. Nowadays, rapid and on-time diagnosis of sicknesses is highly important in that early diagnosis of sickness can be accompanied by more effective cures. Diagnosis of hepatitis is highly difficult as a result of few clinical symptoms in early stages. This study aims to diagnose this sickness using data mining-based methods. Section two studies subject literature .Section three explains Hepatitis database. Section four describes data-mining algorithms used in this article .Final sections provides summary and results.

II. LITERATURE REVIEW

Here are introduced some studies concerning the subject .One of researches in this regard is hepatitis diagnosis by an expert system based on fuzzy facts optimized by a neural network [4] .Another work conductedin this field is decision tree and learning the Law of Dependency to extract rules from hepatitis data in the form of genetic chromosome where the best rules ,those enjoying the highest accuracy, have been extractedusing the power of genetic algorithm optimization and finally, optimized rules have been converted in to a single usable expertise system through Clips programming by expert individuals of this field [2].Another study usesselection-specification method to extract important data features and Support Vector Machine ,SVM, learning method for classification [6].Of conducted researches in this field , method of [8] can be pointed out for extraction features and data classification according to

DOI: 10.15680/IJIRCCE.2016. 0401162

Linear Discriminant Analysis, LDA, and Adaptive Network Fuzzy Inference System, ANFIS [6]. Another method was proposed by [7] according to CART algorithm, C4, 5 algorithm, and ID3 algorithm [7]. invented a method called DIAGFH, a consultation system for rapid diagnosis of hepatitis .In this method, sickness-related information included in a tree-like structure and two conditions indicating whether an individual has hepatitis or not studied the sickness identification [8]. Jilaniet. al., recommended a neural network-based and PCA-based system for Hepatitis C diagnosis .This system has got two stages; initially, feature was extracted by PCA technique and then classification was made using neural network [9]. Uttershwaret. al., stated an expertise system for Hepatitis B diagnosis using logical inference and generalized regression neural networks .Initially, it diagnosed the sickness by an expertise system and then the stage of his sickness was predicted by neural network [10].

III. DATA BASE

A considerable number of medical data are accessible in Website of California University, http://archive.ics.uci.edu. Collection of Hepatitis data includes 155 data samples from two classes, Live and Die. Total number of live samples was 123 cases and the Die ones were 32 cases. Total number of attributes are reported 17 .In addition, another attribute is class label for each sample. Features used from this data collection for classification are listed in table 1.

Table 1: existing features in hepatitis data collection.

Feature
Steroid
Fatigue
Malaise
Anorexia
Liver Big
Liver Firm
Spleen Palpable
Spiders
Ascites
Varices
Bilirubin
Sgot
Albumin

IV. METHODS

The In this study, decision-tree algorithms, neural network, and Support Vector Machine have been used for data classification and forecasting as following:

4.1- Decision Tree:

Decision tree is one of the strong and common tools for classification and prediction. According to training collection, we create one tree .Each internal node of this tree shows one test on a single feature .Each branch reveals the test result, and each leaf keeps label of a class .In this study, Tree C5.0 algorithm has been used to create the tree .Finally, this algorithm produces one decision tree or a collection of rules and regulation. This model classifies fields with the most important information .Each sub-sample is created by the first category from the main one .This trend is repeated until other sub-categories cannot be divided to other or smaller sub-categories .At last, test is taken again at the lowest section of tree and in fact ,importantly diagnosed leaves arepruned and cut [11].

4.2 Neural Network:

Neural network is another strong and common tool for classifying and predicting tool considered as components of modern learning method. Neural networks have been inspired from biological neuron model and they enjoy a substantial number of biological properties of nerves such as being non-linear , simplicity of computational units ,and learning capability .Neural networks are able to realize the relationship of presented information and generalize it to new states (information which is not existing) .Neural network inputs are variables and outputs are cases which need to be predicted or controlled .Artificial neural networks do two main tasks : learning and calling .Learning is weight regulating process of links in one neural network in that a network is able to produce desired output vector as response while receiving stimulant vector by input layer .Calling is acceptance process of an input ,stimulant and production of one output response, stimulation,and production of one output response according to taught weight structure of the network .

4.3 Support Vector Machine:

While trying to discover patterns and classification models, machine learning can be considered as a strong tool.Support Vector Machine is actually one two-class classifier, separating classes by a linear boundary .In this method, the closet samples to decision-making boundary are called Support Vectors .These vectors determine boundary decisionequation .This method shows better performance on data where models are not made with due to using Structural risk minimizing principle implemented by maximizing the distance between two transient clouds from support vector of both classes and despite experimental risk minimizing where it is trying to minimize training error .In this method , linear boundary between two classes are calculated in a way that :

1. All samples in Class +1 are in one side of the boundary and all samples of class -1 are on the other side.
2. Decision boundary is in a way that the nearest distance of training samples in both classes from each other are in perpendicular direction on decision boundary tree.

A considerable number of core functions exist .In this article , radial basis functions and Polynomial functions were applied ,shown in equations (1,2) [12]

$$k(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)(1)$$

$$k(x,y) = (x.y+1)^d \ (2)$$

V. EVALUATION

In this section, test environment is first introduced and then pre-processing operations will be explained before implementation of data-mining algorithms on data collection .Finally, the results of algorithms will be analyzed.

5.1 Operational Environment:

Clementine tool has been used to apply data-mining technique .This tool includes various data-mining algorithms such as, classification, clustering, associative rules, and data pre-processing operation. The difference between this software and other software packages for data processing is applying number of nodes linked with each other within one pattern .In addition, after completion of stages, graphical results can be shown to the user [13].

5.2 Data Pre-Processing:

In this stage, existing records in data collection become statistics of data-mining process .Data collection and pre-processing include selection of data source, elimination of distracted or confusing points, and way of dealing with lost data, and conversion or differencing and reeducation of data.

5.2.1 Lost Data:

Collection of medical data experiences a considerable number of features which have lost values. Total number of 13 features experience lost condition in data collection being used in this research .These features along with their corresponding percentages are listed in table 2.

Table 2: Features with percentage of lost values in hepatitis data collection.

Lost Value %	Feature
0.6	Steroid
0.6	Fatigue
0.6	Malaise
0.6	Anorexia
6	Liver Big
7	Liver Firm
3	Spleen Palpable
3	Spiders
3	Ascites
3	Varices
4	Bilirubin
2.5	Sgot
10	Albumin

The first four policies have been taken in to account to solve the problem of lost values .Each of these policies are explained as following:

First policy: If one feature faces lost values in more than 50 percent of records, then this feature cannot be an effective feature for analyses. Consequently, these kinds of features are eliminated from feature collection.

Second policy : If one feature faces lost values in less than one percent of records , in case it is a numerical feature , mean values of existing ones are substituted with lost values and if it is sequential type ,then it becomesmode of existing valuesin that feature .

Third policy: If one feature experiences lost values in less than 10 percent of records, then that feature is calculated based on type of feature in each class and mean values or class- related mode is substituted considering the record class which is facing lost values.

Fourth policy: If one feature experiences lost values in less than 10 percent of records, then existing classification algorithms are used to estimate lost values.

Now, according to percentages of lost values, appropriate policies have been taken in to account and suitable numbers are listed in table 3 instead of lost values.

Table 3: Features and replaced values for lost values in hepatitis data collection.

Feature	Replaced Values
Steroid	Yes
Fatigue	Yes
Malaise	No
Anorexia	No
Liver Big	Yes
Liver Firm	No
Spleen Palpable	No
Spiders	Live: No, Die: Yes
Ascites	No
Varices	No
Bilirubin	Live:1.14, Die:2.54
Sgot	Live:82.43,Die:99.383
Albumin	Live:3.97,Die:3.15

VI. CREATION OF MODEL

As it was mentioned, data were calculated and analyzed by Clementine software and C5.0, SVN, and Neural network were compared.

Hepatitis analysis modeling using C5.0 algorithm, SVM algorithm, and neural network has been illustrated in Figure1, Figure 2, and Figure 3, respectively.

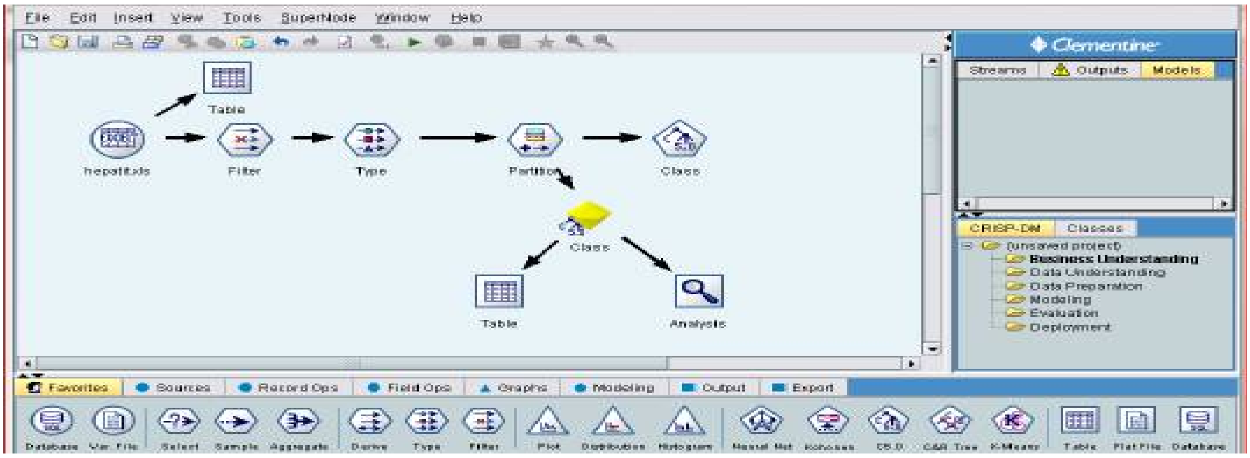
File Edit Insert View Tools SuperNode Window Help
Table
hepatit.xls
Filter
Type
Partition
Class
Class
Table
Analysis
Clementine
Streams Outputs Models
CRISP-DM Classes
(unsaved project)
Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Favorites Sources Record Ops Field Ops Graphs Modeling Output Export
Database Var. File Select Sample Aggregate Derive Type Filter Plot Distribution Histogram Neural Net Kohonen C5.0 C&R Tree K-Means Table Flat File Database

Figure 1: Graphical view of C5.0 model to predict hepatitis in software.

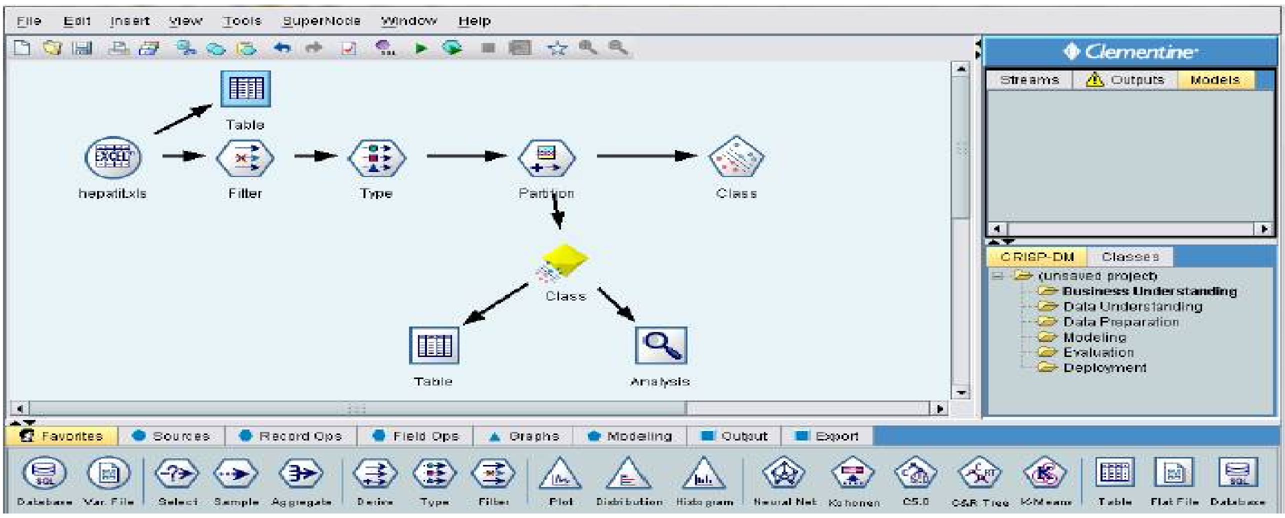
File Edit Insert View Tools SuperNode Window Help
Table
hepatit.xls
Filter
Type
Partition
Class
Class
Table
Analysis
Clementine
Streams Outputs Models
CRISP-DM Classes
(unsaved project)
Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Favorites Sources Record Ops Field Ops Graphs Modeling Output Export
Database Var. File Select Sample Aggregate Derive Type Filter Plot Distribution Histogram Neural Net Kohonen C5.0 C&R Tree K-Means Table Flat File Database

Figure 1: Graphical view of SVM model to predict hepatitis in software.

DOI: 10.15680/IJIRCCE.2016. 0401162

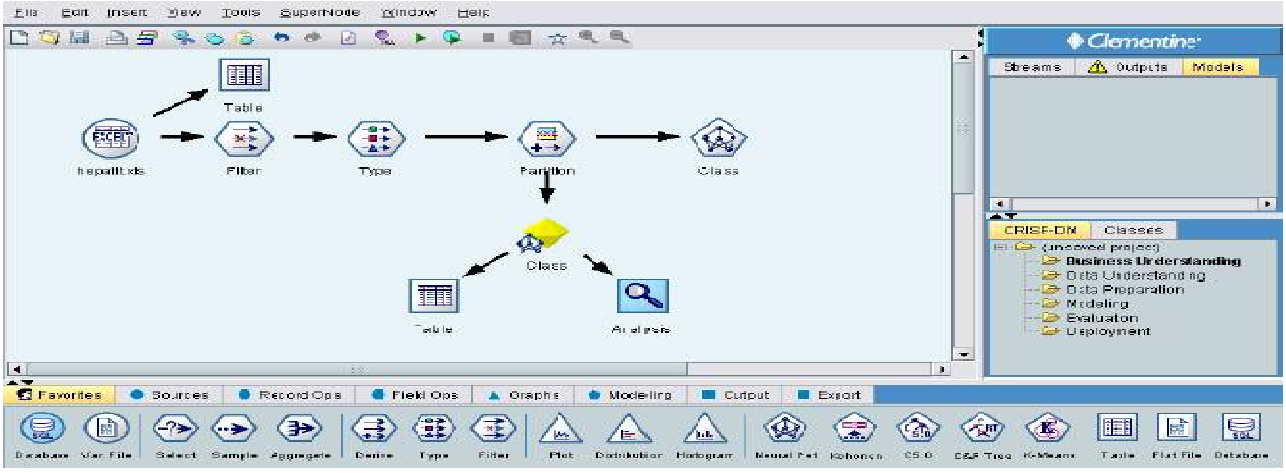

Figure 1: Graphical view of neural network model to predict hepatitis in software.

Data test is done through Partition nodein model .Data have randomly been divided in to two sections :test and train with ratio of 80 percent train and 20 percent test .Various algorithms have been used to predict the fact that an individual has developed hepatitis ,died ,or living .Accuracy obtained from C5.0 algorithm is listed in table 4. Tables 5 and 6 show the accuracies from SVM algorithm function, and finally table 7 shows the accuracy of neural network algorithm.

Table 4: Accuracy of model to predict hepatitis by C5.0 algorithm.

Accuracy	Train		Test	
Correct	109	93.97%	29	74.46%
Incorrect	7	6.03%	10	25.64%
Total	116	100%	39	100%

SVM network structure has two poly functions and RBF, Radial Basis Function .Data were analyzed as above.

Table 5: Accuracy of model to predict hepatitis by SVM algorithm (RB Function).

Accuracy	Train		Test	
Correct	113	97.41%	29	74.46%
Incorrect	3	2.59%	10	25.64%
Total	116	100%	39	100%

Table 6: Accuracy of model to predict hepatitis by SVM algorithm (poly Function).

Accuracy	Train		Test	
Correct	116	100%	29	74.46%
Incorrect	0	0%	10	25.64%
Total	116	100%	39	100%

Table 7: Accuracy of model to predict hepatitis by neural network algorithm

Accuracy	Train		Test	
Correct	99	85.34%	35	89.74%
Incorrect	17	14.66%	4	10.26%
Total	116	100%	39	100%

VII. RESULT ANALYSIS

Table 4 states that software has only been able to find the logic of 109 out of 116 cases it has for training of hepatitis prediction and the remaining seven cases the samples in which software has not been able to find their pattern .Then the software test 39 cases and predict their classes .Out of this number, 29 samples were correctly predicted : it has been able to ,by accuracy of 89.74%, predict the life of patients with hepatitis .Other tables areanalyzed similarlyfor hepatitis prediction. As it can be seen in above tables, the highest accuracy of hepatitis prediction is reported for neural network algorithm .In other words, hepatitis of patients can be predicted by accuracy of 89.74 percent using neural network.

VIII.CONCLUSION

This research reveals that data-mining techniques have offered great promise for discovering hidden patterns, helping clinical specialists for decision making .As it can be seen from above research, accuracy for analysis of different classified data-mining techniques are acceptable to great extent and can help medical specialists in decision making for initial diagnosis and avoiding biopsy. In this research, applying different data-mining techniques on hepatitis data, we intended to discover some information helping physicians to diagnose this sickness.

According to the fact that the more the accessible records are,the better results will be obtained .Using more complete data collection will be beneficial in future researches .Moreover, if collected data is available fordata mining in certain periods,it means that individually related datahave been collected periodically ,following such information and considering the role of time can be taken in to account in future applications .

REFERENCES

1. K Nezhad, MB Minaee, data mining in medicine,third conference of data mining. Tehran (2009).
2. N Arabi, N Rastin, et al. The first national conference of intelligent systems (soft calculations) in science and industries, Islamic Azad University, Ghouchan branch (2013).
3. T Mitchell, Machine Learning. New York: McGraw-Hill (1997).
4. CKemal Polat ,SalihGunes. Medical decision support system based on artificial immune recognition immune system (AIRS), fuzzy weighted pre-processing and feature selection. Science direct Expert Systems with Applications (2007) 484- 490.
5. M Neshat, M Yaghobi, Designing a Fuzzy Expert System of Diagnosing the Hepatitis B Intensity Rate and Comparingit with Adaptive Neural Network Fuzzy System, WCECS (2009) 133-140.
6. G Eason, B Noble, et al. On certain integrals ofLipschitz-Hankel type involving products of Bessel function. Phil. Trans. Roy. Soc. London (1955) 529-551.
7. G Sathyadev, Application of CART algorithm in hepatitis disease diagnosis. Recent Trends in Information Technology (ICRTIT), International Conference on Digital Object Identifier (2011).
8. E Dogantekin, A Dogantekin, Automatic hepatitis diagnosis system based on Linear Discriminant Analysis and Adaptive Network based on Fuzzy Inference System. Expert Systems with Applications (2009) 36.
9. AJ Tahseen, H Yasin, et al. Article: PCA-ANN for Classification of Hepatitis-C Patients. International Journal of Computer Applications, Published by Foundation of Computer Science (2011).
10. GS Uttreshwar, AA Ghatol, Hepatitis B Diagnosis Using Logical Inference and Generalized Regression Neural Networks. Advance Computing Conference.IACC (2009).
11. JP Marques de Sa, Applied statistics: using SPSS, STATISTICA, and MATLAB. Springer(2003).
12. K Rpbert, S Mika "An Introduction of Kernel Based Learning Algorithms". IEEE Transactions on neural Networks (2001);12:181-202.
13. SPSS Inc. Clementine 12.0 Algorithm Guide (1999)105.

DOI: 10.15680/IJIRCCE.2016. 0401162