Comparative Study for Analysis the Prognostic in Hepatitis Data: Data Mining Approach

Fadl Mutaher Ba-Alwi, Houzifa M. Hintaya

Abstract— Data mining techniques are widely used in classification and prediction in the field of bioinformatics to analyze biomedical data. The purpose of the study is to investigate and compare (7) different classification algorithms 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.

Index Terms— data mining algorithms, Hepatitis dataset, Naïve Bayes algorithm, artificial neural network.

1 INTRODUCTION

Artificial Intelligence is one of approach that can train computers to think like human, where it can learn through experience, recognize patterns from large amount of data and also decision making process based from human knowledge and reasoning skills. According to Luger, AI can be defined as the branch of computer science that is concerned with the automation of intelligent behavior. It is combination of science and engineering field in order to make intelligent machines. There are three perspectives in AI; 1) AI can be as a replacement, 2) it can be as assistant and 3) it also can be used to extend human capabilities[1].

Several machine learning techniques or data mining tools like Artificial Neural Networks (ANN) and Rough Sets Theory (RST) are used for data classification. There have been several of research works and surging interests in ANN and developing hybrid system by combining other applications with ANN [2]. The neural network and rough sets methodologies have their place among intelligent classification and decision support systems. Knowledge of the system can be seen as organized data sets with the ability to perform classification.

Nowadays, computers technology and data bases helps human in collecting and storing huge amount of data. The large size of most data bases makes it impossible for human to interpret data. Therefore, computers are needed for extracting new, useful knowledge. Furthermore, other science methods like machine learning, artificial intelligence and logics have made progress and achievements in this field. Today, as we can see the usage of Data Mining and Knowledge Discovery gives more advantages to Statisticians in order to reduce the information stored, to reduce costs, increase sales and revenues, also reduce accidents and failure within data [3] [4].

Classification is the main basic function that can be executed by human brain where the classification phase in data mining human can analyze objects by using some characteristics to find out their similarities and differences. Furthermore, life prognosis of hepatitis is a challenging task in early time because of various interdependent features. Data mining techniques have been extensively used in bioinformatics to analyze biomedical data. Data mining algorithms conducted on several studies and have given efficiently results in prediction and classification of inter-related data. This study is aim to classify and compare the accuracy of hepatitis data base using data mining algorithms approach.

This paper is organized as follows. Section 2 deals with the concept of data mining. Section 3 gives an overview of related work. Section 4 & 5 describes the hepatitis dataset and the overall research process. Section 6 elaborates the classification algorithms and illustrates the classification results. Finally, section 7 discuss and compare among accuracy results.

2 IMPORTANT CONCEPTS:

2.1 Knowledge Data Discovery

Knowledge data discovery (KDD) is a method that used in order to extract useful information from large amount of data in the database[5].

Nowadays, the data mining component of KDD fully relies on known techniques from machine learning, pattern recognition

- *Fadl Mutaher Ba-Alwi is Associate Professor in Artificial Intelligence. Head of Information System Department and Senior Lecturer, Faculty of Computer & IT, Sana'a University, Yemen. E-mail fadl_ba_alwi@hotmail.com*
- *Houzifa M. Hintaya is a junior lecturer at Sana'a University, Faculty of Computer & IT. Yemen. E-mail hozifa99@hotmail.com*

and statistics to find patterns in the data mining steps of KDD process[5]. We can see that KDD is focusing on overall of knowledge discovery from data, including how the data is stored and accessed, how the algorithm can be used or adapted into the large amount of data. The need of KDD is very important because previously in health care industry, specialists or doctors will analyze data time by time[6]. They will analyze current trends of data and after that they will provide report detailing the analysis to support health care organization where the report will be their guides for future decision making and planning for health-care management.

2.2 Hepatitis:

The word hepatitis comes from the Ancient Greek word hepar (root word hepat) meaning 'liver', [7]. In medical, hepatitis means injury to the liver with inflammation of the liver cells. The liver is the largest glandular organ of the body [8]. It weighs about (1. 36 kg). It is reddish brown in color and is divided into four lobes of unequal size and shape.

There are six main hepatitis viruses, referred to as types A, B, C, D, E and G. Hepatitis A and E are typically caused if patients eat the contaminated food or water. Hepatitis B, C and D are typically caused by parental contact by infected body fluid and Hepatitis B also can be infected through sexual contact. It is usually caused by a virus spread by sewage contamination or direct contact with infected body fluids.

2.3 Neural Network

Neural Network (NN), as the name indicates, attempts to mimic the neurological functions of the brain (i.e., neural networks). NN consists of computational nodes that emulate the functions of the neurons in the brain. Each node/neuron as a simple processor is interconnected with other nodes via links with adjustable weights. The link weights are adjusted when the NN is learning or being trained. The nodes are classified into two categories (Input and Output layers) or three categories (Input, Hidden and Output layers as shown in Figure 1).

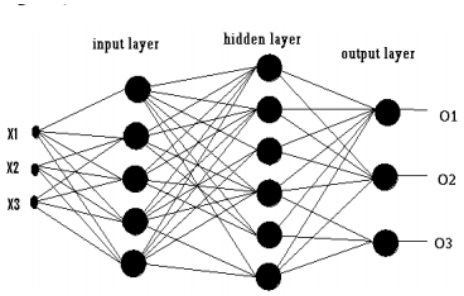

Fig. 1: Multi-layer categories

Recently, neural networks have been used in several pattern classifiers such as speech, medical diagnosis, pattern recogni tion and artificial intelligence applications. It has been the most widely-used classification algorithms because of the availability of high speed computers and large amount of processing power and memory [9]. According to Mavaahebi, neural networks have been used in various fields such as medical and engineering areas [10].

Some characteristics make neural networks used in diagnostic problems. For instance, a set of symptom can be mapped to a set of possible diagnostic classes which known as attribute. In addition, it can reduce the error rate compared to the conventional statistics approaches. Neural network has been proved that it is one of the best practices that can be use in medical diagnosis. The intentions of neural networks evolution techniques especially in medical field are to support specialists or doctors but not to replace them [9],[2].

There are a few numbers of successful applications of neural networks in medical field, for example; myocardial infraction which predicts the heart attack, cancer, pneumonia and brain disorders [2].

2.4 Rough Set Theory

The Rough set theory (RST) has been proposed by Pawlak in 1982. This theory can be use to retrieved or acquitted some data for classification. It also can evaluate the degree of data from database in order to classify the data [11]. RST is more to mathematical approach to imprecision, vagueness and uncertainty, based on the original data sets not any external information. It is suitable for both quantitative, qualitative attributes and discovers hidden facts in data in the form of decision rules. The derived decision rules describe the knowledge contained in the information tables and eliminate the redundancy of original data.

RST also do feature reduction by finding minimal the subsets (reducts) of attributes that are efficient for rule making which is the central part of its process. RS has been applied in medical areas, which are; peritoneal lavage in pancreatitis, toxicity predictions, development of medical expert system rules, prediction of death in pneumonia, identification of patients with chest pain who do not need expensive additional cardiac testing, diagnosing congenital malformations, prediction of relapse in childhood leukemia, and to predict ambulation in people with spinal cord injury [11], [12].

It is proved that RST is a recent intelligent technique that can discover the data dependencies, evaluate the importance of attributes, discover the patterns in dataset, reduce redundancies, and to recognize and classify objects [13].

RST plays a main role in artificial intelligence (AI) and cognitive science, machine learning, knowledge acquisition, decision analysis, knowledge discovery, expert system, decision support systems, inductive reasoning and pattern recognition [12].

2.5 Naïve 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[14]. 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.
Generally, however, the use of the naive Bayesian classifier produces good performance in terms of classification accuracy, despite violations of the attribute independence assumption and is, as such, widely-used in medical data mining . It has also been used as a baseline algorithm for the comparison of other types of classification algorithms.

3. RELATED WORK

Recently, several studies have been conducted and have focused on medical diagnosis. These studies have applied different approaches and have achieved various classification accuracies, usually 75% and higher. Most of the studies dataset has taken from the UCI machine learning repository. Here are some examples:
Robert Detrano's [15] experimental results showed correct classification accuracy of approximately 77% with a logistic-regression-derived discriminant function. The John Gennari's [16] CLASSIT conceptual clustering system achieved 78.9% accuracy on the Cleveland database. L. Ariel [15] used Fuzzy Support Vector Clustering to identify heart disease. This algorithm applied a kernel induced metric to assign each piece of data and experimental results were obtained using a well-known benchmark of heart disease. Ischemic -heart:-disease (IHD) - Support .Vector Machines serve as excellent classifiers and predictors and can do so with high accuracy. In this, tree based: classifier uses nonlinear proximal support vector machines (PSVM).
Polat and Gunes [6] designed an expert system to diagnose the diabetes disease based on principal component analysis. Polat et al. also developed a cascade learning system to diagnose the diabetes. Campos-Delgado et al. developed a fuzzy-based controller that incorporates expert knowledge to regulate the blood glucose level.Magni and Bellazzi devised a stochastic model to extract variability from a self-monitoring blood sugar level time series[17]. Diaconis,P. & Efron,B. (1983) developed an expert system to classify hepatitis of a patient. They used Computer-Intensive Methods in Statistics. Cestnik,G., Konenenko,I, & Bratko,I. designed a Knowledge-Elicitation Tool for Sophisticated Users in the diagnosis of hepatitis.

4. DATA PREPROCESSING

4.1 Dataset

This study conducts experiments on hepatitis dataset. The dataset contains 155 instances distributed between two classes die with 32 instances and live with 123 instances. There are 20 attributes, including the class attribute and 20 missing values. The main goal of the dataset is to forecast the presence or absence of hepatitis virus. This dataset was obtained from UCI machine learning repository.

4.2 Attribute Identification

The last dataset include descriptions of Hepatitis Prognostic Database which can predict either a patient is infected with Hepatitis according to the patient's performance result that can get after few basic physical examinations is done on the patient. This data set can gives a prognosis result either the patient can stay live or die.

Table 1: Data information about Hepatitis diseases

No.	Attribute	Type	Values
1	Class	Categorical	Used as output: - Die - Live
2	Age	Numeric	Numerical values
3	Sex	Categorical	Male, Female
4	Steroid	Categorical	No, Yes
5	Antivirals	Categorical	No, Yes
6	Fatigue	Categorical	No, Yes
7	Malaise	Categorical	No, Yes
8	Anorexia	Categorical	No, Yes
9	Liver Big	Categorical	No, Yes
10	Liver Firm	Categorical	No, Yes
11	Spleen Palpable	Categorical	No, Yes
12	Spiders	Categorical	No, Yes
13	Ascites	Categorical	No, Yes
14	Varices	Categorical	No, Yes
15	Bilirubin	Numeric	0.39, 0.80, 1.20, 2.00, 3.00, 4.00
16	Alk Phosphate	Numeric	33, 80, 120, 160, 200, 250
17	SGOT	Numeric	13, 100, 200, 300, 400, 500
18	Albumin	Numeric	2.1, 3.0, 3.8, 4.5, 5.0, 6.0
19	Protime	Numeric	10, 20, 30, 40, 50, 60, 70, 80, 90
20	Histology	Categorical	No, Yes

4.3 Extraction Hepatitis dataset

The hepatitis dataset contains the screening data of hepatitis disease patients. In the beginning, the dataset was pre-processed to make the mining data process more efficient. In our paper, we used Neural Connection and Weka tools to compare the performance accuracy of data mining algorithms for diagnosis hepatitis disease dataset. Then, the pre-processed dataset is used to remove missing values in order to improve the classification performance. Selection in the tool describes the attribute status of the data present in the hepatitis disease. Using machine learning algorithm such as Naive Bayes, J48, FT Tree and Kstar and then the results are compared. WEKA is a collection of machine learning algorithms for data mining tasks. The algorithms can be applied directly to a dataset. WEKA contains tools for data classification, Associate, clustering and visualization. It is also well suited for developing new machine learning schemes. This paper concentrates on functional algorithms like Naive Bayes, J48, LMT, Neural network, FT Tree and Kstar.

5. Research process

The process involved several stages:

Stage 1 Discretization

The dataset was divided into two parts: training (80%) and testing (20%), in order to guarantee the exactitude of the experimental result and improve the credibility.

Stage 2 pre-processing
Data in the training database must be preprocessed before evaluation. Data preprocessing involved cleaning the data i.e ensuring that the data is free from missing values, noise (contain errors, outlier values) and inconsistencies (discrepancies of units used). Several approaches are available for this purpose. In this study, missing values were replaced with mean value because this method has been commonly used by many researchers [8]. After preprocessing, a complete dataset was obtained and used for experiments

Stage 3 Classifications
In this stage, classification is used to classify data into predefined categorical 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. In our study, classification can be used to help define hepatitis diagnosis and prognosis based on symptoms and health conditions. In this stage, there were 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. In this paper we mainly deal with Naive Bayes, Naive Bayes updatable, FT Tree, KStar, J48, LMT, Neural network.

Stage 4 Accuracy Comparisons and Statistical Results
In this stage, we discussed and compared the accuracy percentage and statistical results among algorithms.

6. Experiments and Results

This paper consists seven different learning algorithms derived from Neural connection and Weka data mining tools, which include naïve bayesian classifier, Naive Bayes updatable, FT Tree, KStar, J48, LMT, and Neural network.
The above algorithms were used to predict the accuracy of Hepatitis dataset. Datasets are divided into training data and test data. Table 2 summarize the best classification accuracy results and Figure 2 is the chart of the results. Obviously, Naive Bayes algorithm has higher classification accuracy than others

Table 2 **Performance study of algorithm**

No.	Algorithm Used	Accuracy %	Time Taken sec
d1	Naive Bayes	96.52	0
2	Naive Bayes updatable	84	00
3	FT Tree	87.10	0.2
4	KStar	83.47	0
5	J48	83	0.03
6	LMT	83.6	0.6
7	Neural network	70.41	-

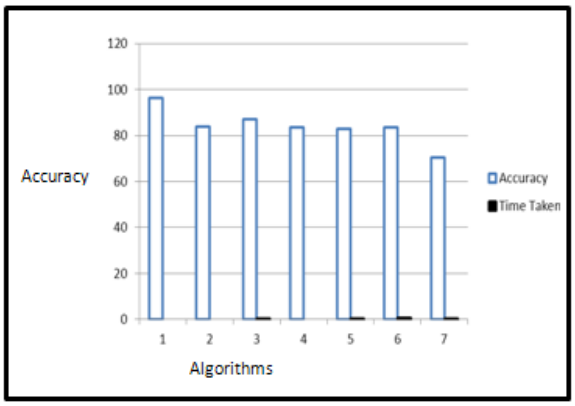

Fig.2 Performance evaluation of the classifiers for Hepatitis data set

Refer to the table 2 above; it is proved that Rough set technique is the best technique to use in analyzing hepatitis data. It gave the highest percentage of accuracy. The best classification algorithm used in Rough set technique is Naïve bayes, which is based on Bayes rule. It is used to reduce complexity in analyzing data.

Through the experiments that have been done to hepatitis datasets, the percentages of accuracy are obtained and the best technique has been determined according to the highest result. This section discussed about the results that have been obtained through the experiments.
Hepatitis datasets have been trained by multi-layer neural network using back-propagation algorithm. There are three units involved in order to execute the neural network which are; input layer, hidden layer and output layer.
Each time data is trained, the training dataset which are 80% from total data is needed. The training dataset was X1 and X2 as shown in the figure 2, in the input signals to responds the output layer of desired output.

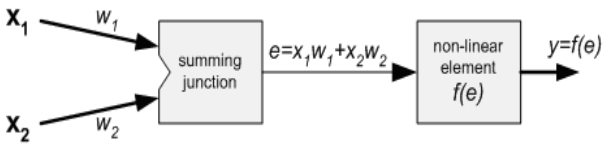

Figure 3: Process of neural network layer

Normalization is done in input layer. It is divided by the value of standard deviation. Number of hidden layer is set. In this experiment, it was set the no hidden layer as one (1) unit. The activation function is selected; the sigmoid and tanh activation function is selected in order to make the comparisons which will gave the best accuracy. The use of activation functions is to introduce the nonlinearity; meaning that without the nonlinearity, the hidden unit would not make the nets powerful to train or execute the data. The weight seed is set in

order to improve the performance of the network. By default, the weight seed value is one (1), but the weight values had been changed until ten (10) to determine which of the weight seed value gave the best result. In the next step, steepest descent learning algorithm is selected. Steepest descent algorithm is popular learning algorithm for back-propagation network. The objective of learning algorithm is to adjust the weight of the network to make sure the output or result will be close as possible to the target. The learning rate, momentum rate and stopping condition values are selected in order to reduce the total error in the network and make the speed of learning more efficient. The results that obtained from the experiment by using the Neural Connection were low comparing to other algorithms used in Weka.

The second technique that is used in this experiment is rough set theory, by using Weka to analyze the Hepatitis dataset. Data was trained and it has been divided into two; training and testing. After that, those data have been discretized in order to group the data which have continuous values in the attributes. There are few numbers of discretization processes, which are; Boolean reasoning, Entropy, Naïve and Semi Naïve. After discretization process, Generation rules process i.e reduction is used. Reduction is techniques which will eliminates the unused attributes and create the minimal subset of attributes for decision table.

Refer to the table II; it is proved that Rough set technique is the best technique to use in analyzing hepatitis data. It gave the highest percentage of accuracy. The best classification algorithm used in Rough set technique is Naïve bayes, which is based on Bayes rule.

7. Conclusion and Future Work

Improving accuracies of machine-learning algorithms is vital in designing high performance computer-aided diagnosis systems. The present study was carried out to justify the performance of ensemble methods in medical data set that can be used for making effective diagnosis, which in turn would increase the health index. In this paper, several algorithms were used. It comes up with a considerable result which leads to the success of the research and to achieve its main goal by comparing different algorithms.

In addition, this paper has proved that Rough set technique is better compared to Neural Network especially in analyzing medical data. The prediction of the outcome is more specific and accurate using Rough set technique. It helps many specialists and doctors in predicting and diagnosing patients.

This paper deals with the results in the field of data classification obtained with Naive Bayes algorithm, Naive Bayes updatable algorithm, FT Tree algorithm, KStar algorithm, J48 algorithm, LMT algorithm and Neural network.

On the whole performance made known Naive Bayes algorithm when tested on hepatitis datasets, time taken to run the data for result is fast when compared to other algorithms. It shows the enhanced performance according to its attribute. Attributes are fully classified by this algorithm and it gives 96.52% of accurate result. Based on the experimental results the classification accuracy is found to be better using Naive Bayes algorithm compared to other algorithms. From the above results Naive Bayes algorithm plays a key role in shaping improved classification accuracy of hepatitis dataset.

In future, it is possible to extend the research by using different classification techniques and association rule mining for large number of patients. Moreover, it is necessary to apply fuzzy learning models for further enhanced fore casting of hepatitis virus.

References

[1] H. Kaur, "Artificial Intelligence: Bringing expert knowledge to computers," *Discovery*, vol. 2, 2012.

[2] E. Grossi, *Artificial Neural Networks - Methodological Advances and Biomedical Applications* InTech, 2011.

[3] A. Almonayyes, "Multiple explanations driven na¨ive bayes classifier.," *UCS*, vol. 12, pp. 127-139, 2006.

[4] J. D. M. Rennie, "Improving multi-class text classification with naive bayes," *MIT AI-TR*, 2001.

[5] F. U., "From Data Mining to Knowledge Discovery in Databases," *American Association for Artificial Intelligence*, 1996.

[6] K. Polat, "An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease," *Dig. Signal Process*, vol. 17, pp. 702-710, 2007.

[7] T. Karthikeyan, "Analysis of Classification Algorithms Applied to Hepatitis Patients," *International Journal of Computer Applications*, vol. 62, 2013.

[8] P.Rajeswari, "Analysis of Liver Disorder Using Data mining Algorithm " *Global Journal of Computer Science and Technology*, vol. 10, 2010.

[9] F. Amato, "Artificial neural networks in medical diagnosis " *Journal of APPLIED BIOMEDICINE*, vol. II, pp. 47-58, 2013.

[10] M. Mavaahebi, "A Neural Network and Expert Systems Based Model for Measuring Business Effectiveness of Information Technology Investment," *American Journal of Industrial and Business Management*, vol. 3, 2013.

[11] A. Vit´oria, "From Rough Sets to Rough Knowledge Bases," *Fundamenta Informaticae*, pp. 1-32, 2003.

[12] Z. Pawlak, "Rough sets: Some extensions," *Information Sciences*, vol. 177, pp. 28-40, 2007.

[13] Hassanien, "Rough Sets Data Analysis in Knowledge Discovery: A Case of Kuwaiti Diabetic Children Patients," *Advances in Fuzzy Systems*, p. 13, 2008.

[14] S. B. Kotsiantis, "Increasing the Classification Accuracy of Simple Bayesian Classifier," *AIMSA*, pp. 198-207, 2004.

[15] R. Detrano, "International application of a new probability algorithm for the diagnosis of coronary artery disease," *American Journal of Cardiology*, vol. 64, pp. 304-310, 1989.

[16] G. John, "Models if incremental concept formation," *Journal of Atificial Intelligence*, pp. 11-61, 1989.

[17] P. Magni, "A stochastic model to assess the variability of blood glucose time series in diabetic patients self-monitoring," *IEEE Trans. Biomed. Eng*, vol. 53, pp. 977-985, 2006.

[18] W. X, "Top 10 algorithms in data mining," *Knowl Inf Syst*, vol. 14, 2008.

[19] Z. Pawlak, "ROUGH SETS PRESENT STATE AND FURTHER

PROSPECTS," *intelligent Automation and Soft Computing*, vol. 2, pp. 95-102, 1996.

[20] G. F. Luger, *Artificial intelligence: structures and strategies for complex problem solving*, Fifth Edition ed. New York.: Addison-Wesley, 2005.

[21] A. Hassanien, "Rough Sets Data Analysis in Knowledge Discovery: A Case of Kuwaiti Diabetic Children Patients," *Advances in Fuzzy Systems*, vol. 2008, p. 13, 2008.

[22] A. L. Gamboa, "Hybrid Fuzzy-SV Clustering for Heart Disease Identification," *CIMCA-IAWTIC*, vol. 6, 2006.

[23] S. G.P, "The KDD Process for Extracting Useful Knowledge from Volumes of Data," *COMMUNICATIONS OF THE ACM*, vol. 39, 1996.

[24] D. Dumitru, "Prediction of recurrent events in breast cancer using the Naive Bayesian classification," *Annals of University of Craiova, Math. Comp. Sci. Ser*, vol. 36, pp. 92-96, 2009.