Analysis of Classification Algorithms Applied to Hepatitis Patients

T.Karthikeyan, PhD.
Associate Professor
P.S.G. College of Arts and Science
Coimbatore, India.

P.Thangaraju
Research Scholar, Bharathiar University
Asst. Professor, Bishop Heber College
Tiruchirappalli, India.

ABSTRACT

This paper mainly deals with various classification algorithms namely, Bayes.NaiveBayes, Bayes.BayesNet, Bayes. NaiveBayesUpdatable, J48, Randomforest, and Multi Layer Perceptron. It analyzes the hepatitis patients from the UC Irvine machine learning repository. The results of the classification model are accuracy and time. Finally, it concludes that the Naive Bayes performance is better than other classification techniques for hepatitis patients.

General Terms

Data mining, classification, hepatitis

Keywords

Naive bayes, Multi Layer Perceptron, Random Forest, J48.

1. INTRODUCTION

The word hepatitis comes from the Ancient Greek word hepar (root word hepat) meaning 'liver', and the Latin it is meaning inflammation. Hepatitis means injury to the liver with inflammation of the liver cells. The liver is the largest gland in the human body. It weighs approximately 3 lb (1.36 kg). It is reddish brown in color and is divided into four lobes of different sizes and lengths. It is also the largest internal organ (the largest organ is the skin). It is below the diaphragm on the right in the thoracic region of the abdomen. Blood reaches the liver through the hepatic artery and the portal vein. The portal vein carries blood containing digested food from the small intestine, while the hepatic artery carries oxygen-rich blood from the aorta. The liver is made up of thousands of lobules; each lobule consists of many hepatic cells. Hepatic cells are the basic metabolic cells of the liver. The liver has a wide range of functions, including:

- Detoxification (filters harmful substances form the blood, such as alcohol)
- Stores vitamins A, D, K and B12 (also stores minerals)
- Protein synthesis
- The production of biochemicals needed for digestion, such as bile
- Maintains proper levels of glucose in the blood
- Produces 80% of your body's cholesterol
- The storage glycogen
- Decomposing red blood cells
- Synthesizing plasma protein
- The production of hormones
- Produces urea

Most liver damage is caused by 3 hepatitis viruses, called hepatitis A, B and C. However, hepatitis can also be caused by alcohol and some other toxins and infections, as well as from our own autoimmune process. About 250 million people globally are thought to be affected by hepatitis C, while 300 million people are thought to be carriers of hepatitis B. Not all forms of hepatitis are infectious. Alcohol, medicines, and chemical may be bad for the liver and cause inflammation. A person may have a genetic problem, a metabolic disorder, or an immune related injury. Obesity can be a cause of liver damage which can lead to inflammation. These are known as non-infectious, because they cannot spread form person-to-person.

Life prognosis of hepatitis is a challenging task in early stage due to various interdependent features. A model can be developed which can used in life prognosis of hepatitis diseases. Data mining techniques have been extensively used in bioinformatics to analyze biomedical data. Data mining algorithms can be used efficiently in prediction and classification of inter-related data. The objective of this analysis is classify and scaling the accuracy of hepatitis data base. WEKA (Waikaato Environment for Knowledge Acquition) is the most widely used data mining tool which support huge amount of data mining algorithm for classification.

This paper is organized as follows. The section 2 deals with the concept of data mining and describes the overall research process. The section 3 elaborates with classification algorithms like decision trees, naive bayes and neural networks. The section 4 discuss with the data processing and attributes information. The section 5 illustrates the classification results.

2. DATA MINING

Data Mining is defined in many ways in different situations. Major definitions used in literature are refers to the finding of relevant and useful information from databases [1, 2] and also deals with finding of patterns and hidden information from a large database [3].Data Mining is also known as Knowledge Discovery in Databases (KDD) which is defined as the non–trivial extraction of implicit, previously unknown and potentially useful information from the data [4]. The term Knowledge Discovery in Databases or KDD refers to the broad process of finding knowledge in data and emphasizes the "high-level" application of particular data mining methods. It is of interest to researchers in machine learning, pattern recognition, databases, statistics, artificial intelligence, knowledge acquisition for expert systems and data visualization. The unifying goal of the KDD process is to extract knowledge from data in the context of large databases [2, 4]. Using data mining methods or algorithms, the techniques which extract and identify the deemed knowledge, dealing to the specifications of measures and thresholds, using a database along with any required preprocessing, sub sampling, and transformations of that database. An outline of the steps of the KDD Process and the overall process of finding and interpreting patterns from data involves the repeated application of the following steps are developing an understanding of the application domain and the relevant prior knowledge and identifying the goal of the process from the customer’s viewpoint. Creating a target data set which performs selecting a data set or focusing on a subset of variables or data samples on which discovery is to be performed. Data cleaning and preprocessing includes removal of noise or outliers for collecting necessary information to model or account for noise and deciding on strategies for handling missing data fields and according for time sequence information and known changes. Data reduction and projection consist of finding useful features to represent the data depending on the goal of the task. Using dimensionality reduction or transformation methods, the method has to reduce the effective number of variables or to find invariant representations for the data. The data mining task consist of the KDD process such as classification, regression, clustering and so on [2, 4]. Exploratory analysis, model and hypothesis selection to be composed of the data mining algorithms and the selecting methods to be used for searching for patterns in the data, deciding which models and parameters may be appropriate and matching a particular data mining method with the overall criteria of the KDD process.

The Figure 1 shows the overall research process. The data set is taken the UCI machine learning repository and the data undergone a preprocessing and the classification techniques are applied to the data set. The classification accuracy and time to build the model is tabulated and finally research conclusion is derived.

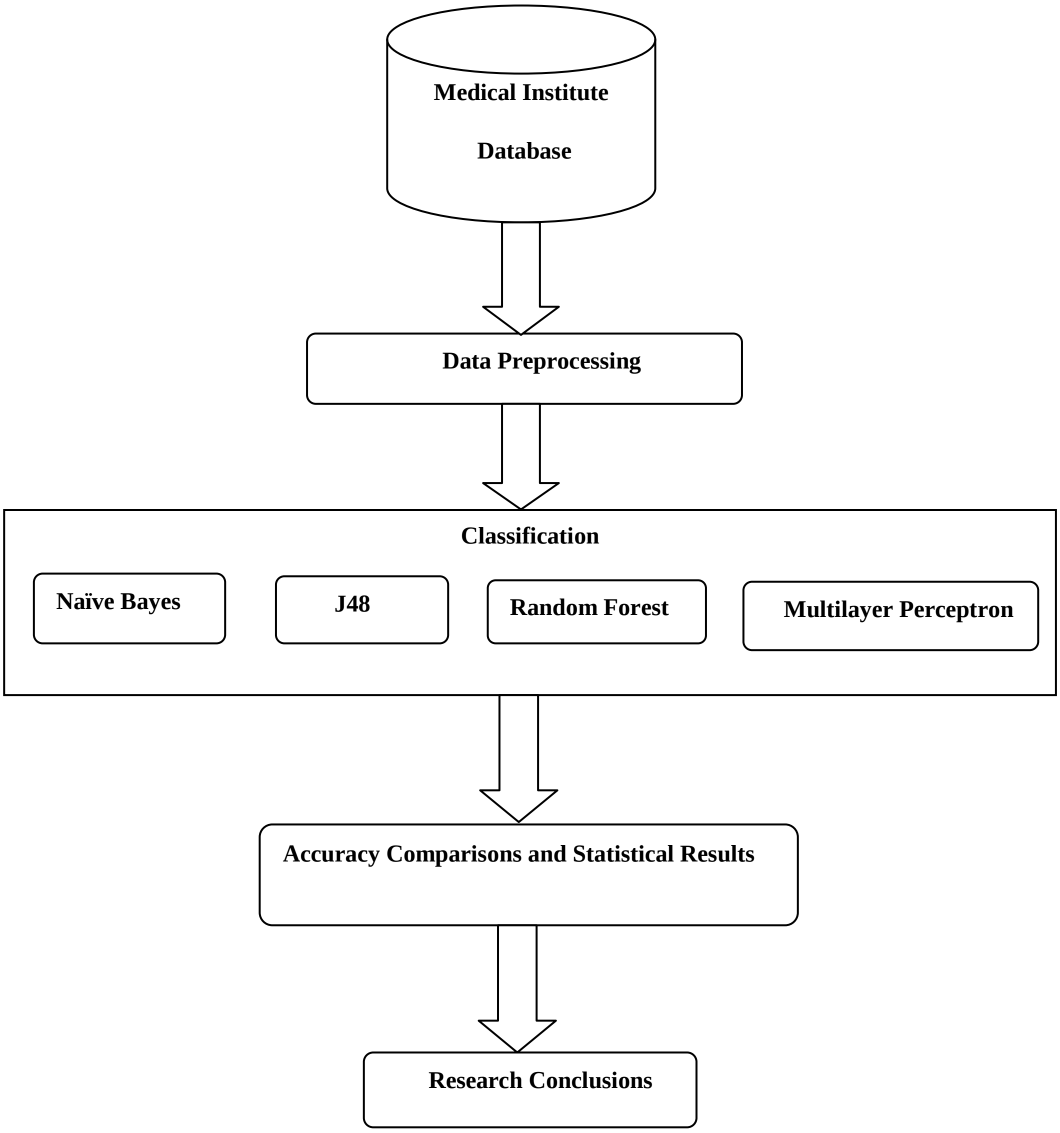

Figure 1: Research Process

3. CLASSIFICATION

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. For example, banks have constructed classification models to categorize bank loan and mortgage applications into risky or safe. In the medical field, classification can be used to help define medical diagnosis and prognosis based on symptoms and health conditions. Classification is a two-step process consisting of training and testing. The first step, training, builds a classification model, consisting of classifying rules,

by analyzing training data containing class labels (an example of A classification rule is "IF Lung_Cancer Family_History = "yes" AND Smoking= "yes" THEN Scan=required"). Some classifiers like SVM use mathematical formula rather than IF-THEN rules for better accuracy. Classifying rules are not necessarily 100% true; generally, rules with 90–95% accuracy are regarded as solid rules.

The accuracy of a classifier depends on the degree to which classifying rules are true. The second step, testing, 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. The testing process is very simple and computationally inexpensive as compared to the training step, which is complex and requires substantial computational resources. This paper mainly deals with decision tree, naïve bayesian classifier and neural networks.

3.1 Decision Trees

Ross Quinlan introduced a decision tree algorithm (known as Iterative Dichotomiser (ID)3) in 1979 [5]. C4.5 [6], as a successor of ID3, is the most widely-used decision tree algorithm. Decision tree classifiers construct a flowchart-like tree structure, (as shown in Figure 2) in a top down, recursive, divide-and-conquer, manner. The Attribute Selection Method (ASM) is the key process in the construction of a decision tree. The ASM selects a splitting criterion that best splits the given records into each of the class attribute whose sorting result is closest to the pure partitions by the class in terms of class values. Selected attributes become nodes in a decision tree. For example (as seen in Figure 2 ASCITES is the first attribute (or the root node) selected by ASM. Decision trees which consist of IF-THEN rules are classification models. In other words, constructing a decision tree is the training step of classification. The major advantage to the use of decision trees is the class-focused visualization of data. This visualization is useful in that it allows users to readily understand the overall structure of data in terms of which attribute mostly affects the class (the root node is always the most significant attribute to the class). A drawback to this method occurs when a data set contains many attributes. In this case, the decision tree may be too complex to be easily understood. To resolve the problem, tree pruning approaches are applied to such decision trees. These approaches resolve the problem of over fitting the dataset by using statistical methods to prune the least important branches so that users can readily capture the overall structure of data. Pruning is a useful technology even though there may be are some minor errors in the trees generated under this method.

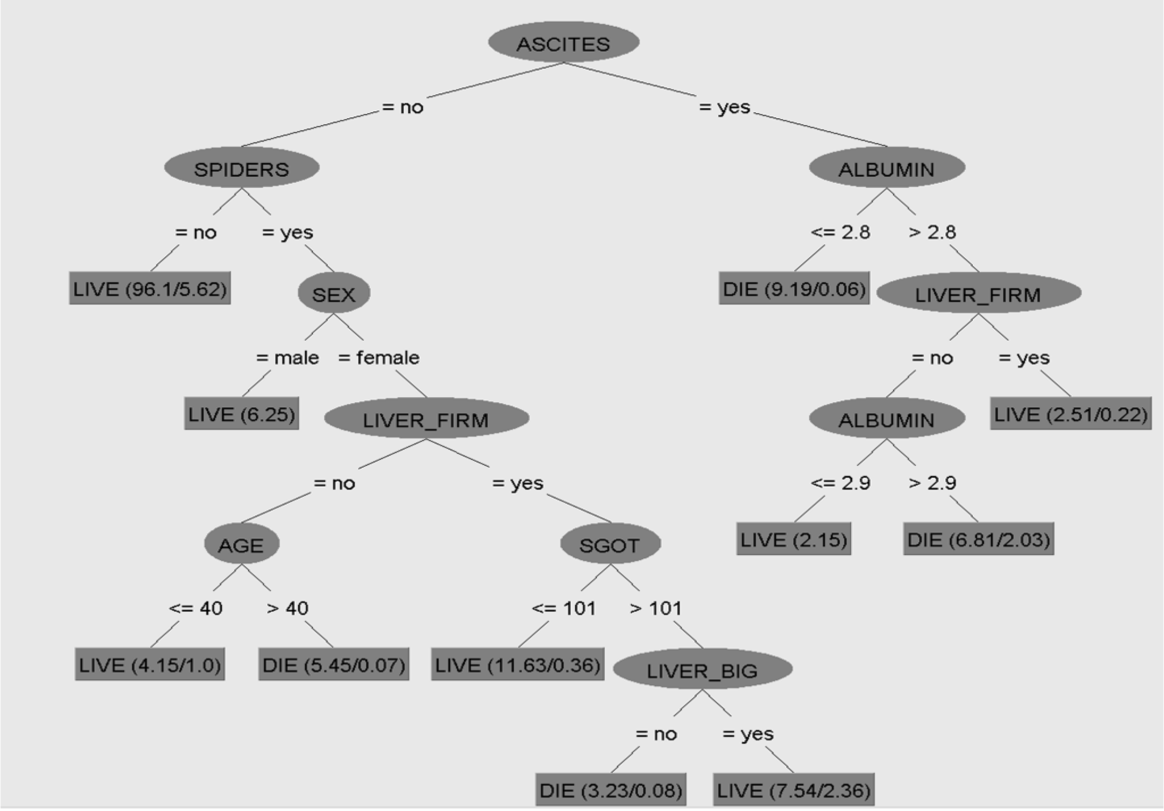

Figure 2: Decision Tree for Hepatitis Data Set

3.2 Naïve Bayesian Classifier

A Naive Bayesian classifier based on Bayes theorem is a probabilistic statistical classifier. Here, the term "naive" indicates conditional independence among features or attributes. The "naive" assumption greatly reduces computation complexity to a simple multiplication of probabilities. The major advantage of the Naive Bayesian classifier is its rapidity of use. This rapidity occurs because it is the simplest algorithm among classification algorithms. 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. A major drawback of this algorithm is its fundamental assumption that all attributes are independent one another. In many cases this assumption is unrealistic. For example, in the medical field, many patient symptoms and health conditions are strongly related each other (e.g., blood pressure and body mass index (BMI)), which may result in some aberration in the resulting classification. 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 [6, 7]. It has also been used as a baseline algorithm for the comparison of other types of classification algorithms.

3.3 Neural Networks

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 3).

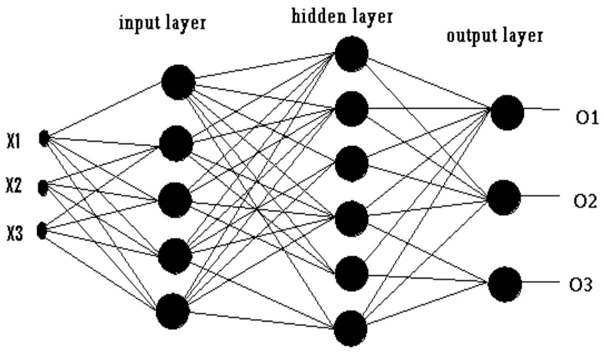

Figure 3: Multi layer Perceptron

However, the input layer is not considered or counted when neural network algorithms are classified because the input layer is simply used to take the input values and pass them to the next layer. For example, a neural network algorithm having input, hidden, and output layers is called a two (or multi)-layer neural network [8, 9]. The most widely used NN is multi-layer perceptron with back-propagation, which is available in WEKA because its performance is considered superior to other NN algorithms including self-organizing map [10]. NN was developed in the early 20th century. It was regarded as the best classification algorithm until the introduction of the decision trees and the Support Vector Machine (SVM). This regard is one of the reasons why NN has been the most widely-used classification algorithm in various biomedicine and healthcare fields [11, 12]. NN has several disadvantages. First, NN requires many parameters, including the optimum number of hidden layer nodes that are empirically determined, and its classification performance is very sensitive to the parameters selected [13]. Second, its training or learning process is very slow and computationally very expensive.

4. DATA PREPROCESSING

Dataset used in this model should be more precise and accurate in order to improve the predictive accuracy of data mining algorithms. Dataset which is collected may have missing (or) irrelevant attributes. These are to be handled efficiently to obtain the optimal outcome from the data mining process

4.1 Attribute Identification

Dataset collected from UC Irvine machine learning repository which consists of 155 instances and 19 attributes with the class stating the life prognosis yes (or) no. The dataset consist of 14 nominal attribute and 6 multi-valued attributes. The attributes which are identified are shown in Table 1, which consist of the class, age, sex, steroid, antivirals, fatigue, malaise, anorexia, liver big, liver firm, spleen palpable, spiders, ascites, varices, bilirubin, alk phosphate, SGOT, albumin, protime, histology.

Table 1. Attribute Details of the Patients

Attributes	value
Class	die (1), live (2)
Age	numerical value
Sex	male (1), female (2)
Steroid	no (1), yes (2)
Antivirals	no (1), yes (2)
Fatigue	no (1), yes (2)
Malaise	no (1), yes (2)
Anorexia	no (1), yes (2)
Liver Big	no (1), yes (2)
Liver Firm	no (1), yes (2)
Spleen Palpable	no (1), yes (2)
Spiders	no (1), yes (2)
Ascites	no (1), yes (2)
Varices	no (1), yes (2)
Bilirubin	0.39, 0.80, 1.20, 2.00, 3.00, 4.00
Alk Phosphate	33, 80, 120, 160, 200, 250
SGOT	13, 100, 200, 300, 400, 500
Albumin	2.1, 3.0, 3.8, 4.5, 5.0, 6.0
Protime	10, 20, 30, 40, 50, 60, 70, 80, 90
Histology	no (1), yes (2)

5. RESULTS

The analysis and interpretation of classification is time consuming process that requires a deep understanding of statistics. The process requires a large amount of time to complete and expert analysis to examine any classification and relationships within the data.

5.1 Data mining Results

The WEKA is an open source collection of machine learning algorithms and data processing tools. WEKA data mining software is used to determine if any advantage could be gained in both time saving and interpretation of the hepatitis data set. The application of the data to WEKA required that some preprocessing be undertaken. The data set produced in excel for the statistical processes were copied and then converted to CSV (Comma Separated Values) file format to allow them to be applied to WEKA. The CSV file extension allowed initial analysis to be conducted with later conversion to be taken into an ARFF (Attribute-Relation File Format) WEKA data file for the experimental outcome to be saved.

The data mining platform allowed number of data interpretations including classify, cluster associate routines to be conducted after the preprocessing stage. The hepatitis data set did not require any filtering because of the limited amount of missing values and the outcomes required by the researchers. The initial screen provides a set of information that is required by the researchers and took a large amount of time to complete with the current statistical methods.

The full hepatitis data set was applied to the Naive Bayes to classify the hepatitis patients and could be established with the model being constructed using a training model to classify the training data set and see the correctly classified instances and also apply the Naive Bayes to test set and see the correctly and incorrectly instances. Determine the accuracy when compared with each other.

The results are when naive bayes classifier is applied to hepatitis data set the instances are better than other classifiers. The other classifiers like J48, trees, Randam Forest is also applied to the hepatitis data and the results are shown in the Table 2.

Table 2. Classifiers Statistical Results

Classifier	Mean Absolute Error	Root Mean Square Error	Kappa Statistics
Bayes.NaiveBayes	0.1594	0.3375	0.5451
Bayes.BayesNet	0.1609	0.3556	0.3852
Bayes.NaiveBayes Updatable	0.1594	0.3375	0.5451
J48	0.2251	0.3288	0.4536
Random Forest	0.2791	0.3471	0.3051
Multilayer Perceptron	0.1884	0.4043	0.4682

5.2 Experimental Results

The time build the Naïve Bayes classifier is less than the remaining classifier is shown in Table 3. Kappa statistics is a measure of the degree of non random agreement between observers and/or measurement of a specific categorical variable. The root mean square error and mean absolute error of Byes.NaiveBayes are also minimum with compare to other classifiers. So the Naive Bayes classifier is the efficient classification technique among remaining classification technique.

Table 3. Experimental Results for Time and Accuracy

Classifier	Time Taken	Accuracy
Bayes.NaiveBayes	0.00	84%
Bayes.BayesNet	0.00	81%
Bayes.NaiveBayes updatable	0.00	84%
J48	0.03	83%
Random forest	0.05	83%
Multilayer perceptron	17.94	83%

5.3 Graph Results

The Figure 4 shows the graphical representation of time and accuracy results of prediction of hepatitis patients based on hepatitis dataset. It clearly depicts that naïve bayes performs better in accuracy in less time.

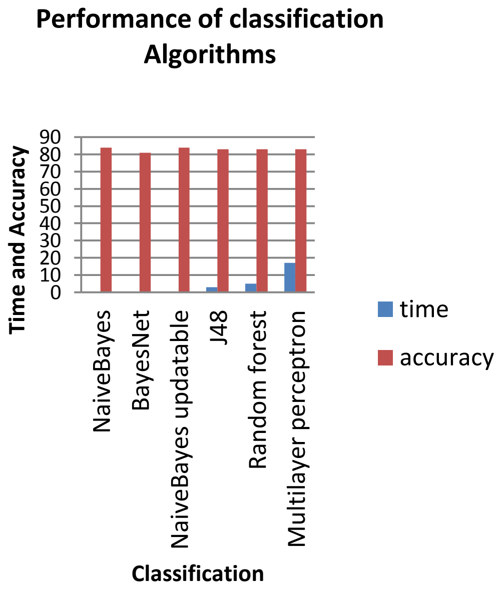

Figure 4: Graphical Representation of Naive Bayes with Existing Techniques

6. CONCLUSION

The experiment conducted and analyzed small number of traits contained within the data set to determine their classification accuracies and calculate the time to build the model and found some statistics. The hepatitis data profiles that are used in this research were selected for completeness and for each classification of hepatitis patients. The result of the experiment proves that Naïve Bayes consumes less time in detection hepatitis patients data set using WEKA tool. Moreover the limitation is that it considers only small traits to detect the virus. Complex terminology is required to predict the results more accuracy. The recommendations arises from this research implies the data mining techniques may be applied in the field of medical research in future as they will provide research tools for comparison of large amount of data.

In future, it is possible to extend the research by using different clustering 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.

7. REFERENCES

[1] Arun K. Pujari, Data Mining Techniques, *Universities Press (India) Ltd,* 2001.

[2] Jiawei Han, Micheline Kamber, Data Mining Concepts and Techniques, *Elsevier.*

[3] Klosgen W, Zytkow JM, Handbook of Data mining andKnowledge Discovery, *Oxford University Press*, 2002.

[4] M.S.Chen, J.hans, P.SYu, Data mining: A overview from a data base perspective, *IEEE transaction on Knowledge and data engineering* 8(6), pp. 866-883, 1996.

[5] Quinlan, J. R., C4.5: programs for machine learning. Morgan Kaufmann, Amsterdam, 1993.

[6] Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, MotodaH, McLachlan GJ, Ng A, Liu B, Yu PS, Zhou Z-H, Steinbach M, Hand DJ, Steinberg D , Top 10 algorithms in data mining. *Knowl Inf Syst* 14, pp.1–37, 2008.

[7] Diana Dumitru , Prediction of recurrent events in breast cancer using the Naive Bayesian classification, *Annals of University of Craiova, Math. Comp. Sci. Ser.* Volume 36(2), 2009.

[8] Nurnberger A, Pedrycz W, Kruse R , Neural networkapproaches. In: Klosgen W, Zytkow JM (eds) Handbook of data mining and knowledge discovery. *Oxford University Press*, 2002.

[9] Hammerstrom D, Neural networks at work. *IEEE Spectr:*pp.26–32 (June), 1993.

[10] Delen, D., Walker, G., and Kadam, A., Predicting breast cancer survivability: a comparison of three data mining methods. *Artif. Intell. Med.* 34, pp.113–127,2005.

[11] Kaur, H., and Wasan, S. K., Empirical study onapplications of data mining techniques in healthcare. *J. Comput. Sci.* 2(2), pp. 194–200, 2006.

[12] Ubeyli, E. D., Comparison of different classification algorithms in clinical decision making. *Expert syst* 24(1), pp. 17–31, 2007.

[13] Schwarzer, G., Vach, W., and Schumacher, M., On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology. *Stat. Med.* 19, pp. 541–561, 2000.