ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 6, Issue 1, January 2017

Analysis of feature Selection and Classification algorithms on Hepatitis Data

Nancy.P, Sudha.V, Akiladevi.R

***Abstract*— Data Mining is becoming one of the leading techniques applicable in a variety of area. One such area is predictive analysis in medical field. In this paper we investigate the performance of about 15 data mining classification algorithms viz. Rnd Tree, Quinlan decision tree algorithm (C4.5), K-Nearest Neighbor algorithm etc., on a large dataset from the "Hepatitis dataset" (derived from the UCI Machine Learning Repository) that comprises of 20 attributes (including class) and 155 instances. Also we investigate on the importance of feature selection and applied three feature selection algorithms namely Fisher filtering, Relief filtering, Step Disc and classified the dataset using 15 most common classifiers. The results of this study indicate the level of accuracy as well as the importance of all the instances in detecting the survival of a person in future. The classification algorithms BVM,CVM and Rnd Tree produced 100 percent accuracy for classification of all the training data under bivalued classes. The study also revealed that the feature selection algorithms as mentioned above are not suitable for this dataset for effective classification. The classification algorithm was also applied to verify it's correctness in classifying test data.**

***Index Terms*— Data mining, classification, hepatitis, Naive bayes, Multi Layer Perceptron, Random Forest, J48.**

I. 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.

***Nancy.P**, Computer Science, Rajalakshmi Engineering College Chennai, India, 9786675376*
***Sudha.V**, Computer Science, Rajalakshmi Engineering College, Chennai, India, 9952005673.*
***Akiladevi.R**, Computer Science, Rajalakshmi Engineering College, Chennai, India, 7708807118.*

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 [1]

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. Tanagra 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 that describes the overall research process and the related works. The section 3 elaborates the details of the dataset, feature selection algorithms and classification algorithms like decision trees, naive bayes and neural networks. The section 4 describes the experimental results and discussion.

II. DATA MINING PROCESS AND RELATED WORK

A. 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 [2]. 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, 3]. 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. 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 in terms of error rate is tabulated and finally research conclusion is derived on.

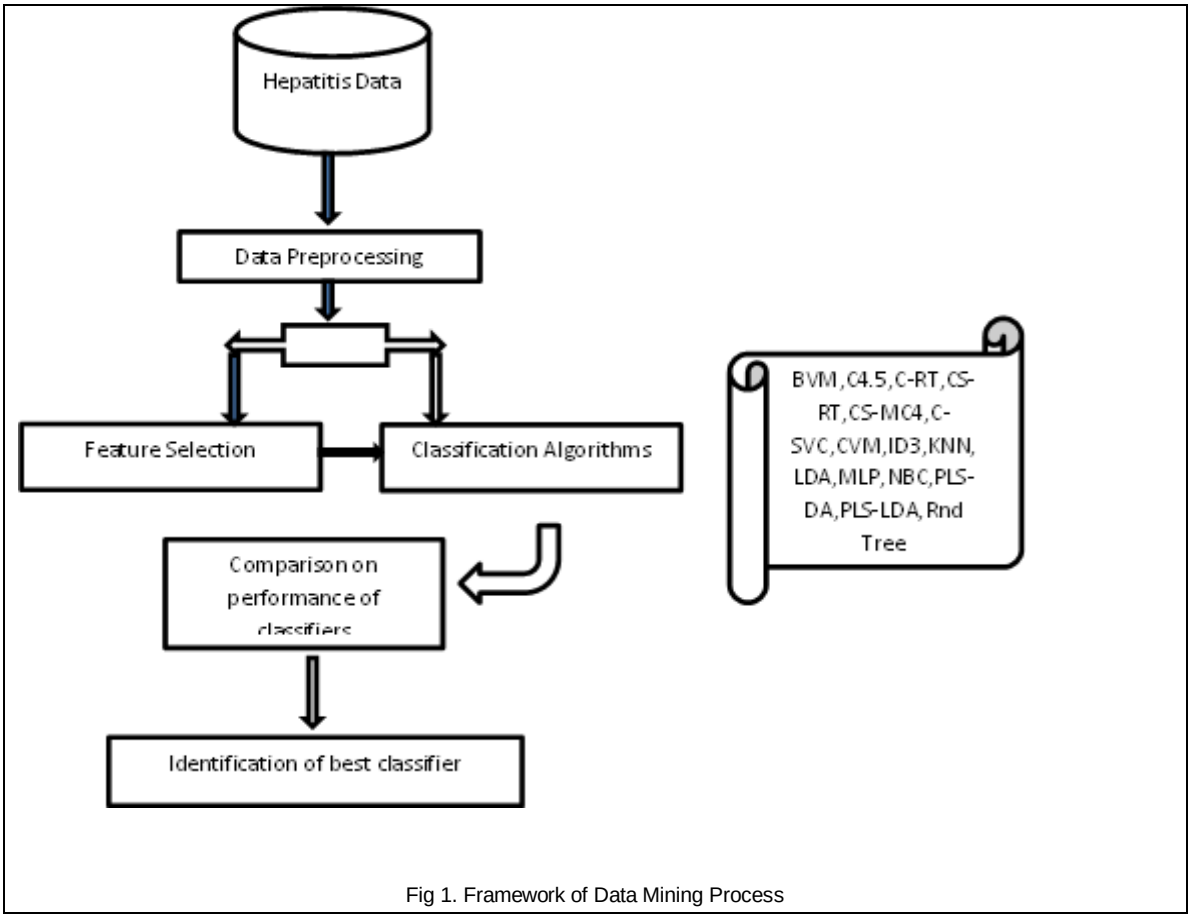

Fig 1. Framework of Data Mining Process

YYılmaz Kayaa et al. [4] developed a new system for diagnosis of hepatitis disease. RS-ELM works in two stages where redundant features are removed in first stageand classification is applied to remaining features in the second stage. The classification accuracy was about 96.49% using RS-ELM model.

Javed Salimi Sartakhti et al. [5] presented a novel machine learning method using hybridized Support Vector machine and simulated annealing for hepatitis diagnosis. It is a stochastic method used for difficult optimization problems.

Duygu et al. [6] proposed an intelligent hepatitis diagnosis system using Principle Component Analysis and Least Square Support Vector Machine Classifier (PCA–LSSVM) which also emphasized on feature extraction and reduction for efficient classification.

G. Sathya Devi [7] proposed the application of CART algorithm in Hepatitis Disease Diagnosis using decision trees C4.5 algorithm, ID3 algorithm and CART algorithms and obtained a classification accuracy of 83.2%.

A.H.Roslina et al. [8] implemented a prediction of hepatitis prognosis using Support Vector Machines and Wrapper Method. To remove the noise features wrapper methods were used before classification. Support Vector Machines showed the accuracy in enforcing feature selection first. Features selection were implemented to minimize noisy or irrelevance data. The accuracy rate was increased concurrently the clinical lab test cost and time was reduced. This was achieved by combining Wrappers Method and SVM techniques.

Fadl Mutaher et al. [9] presented the comparative analysis in the prognostic of hepatitis data using Rough set technique It was found that the performance and time taken to run the hepatitis data is fast in Naive Bayes algorithm. The results obtained were compared with other algorithms like, Naive Bayes up-datable algorithm, FT Tree algorithm, Kstar algorithm, J48 algorithm, LMT algorithm and neural network. Attributes were fully classified and the result obtained was of 96.52%.

III. DATASET AND ALGORITHMS

The data set Hepatitis is taken from UCI repository. The details of the data set is given in Table 1

Table 1
Hepatitis data set

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

A. *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.

B. *Feature Selection*

After replacing the missing values, some preprocessing of the data is to be carried out to proceed further. Feature Reduction is one of the preprocessing techniques. In this phase the important features required to implement the Classification Algorithm are identified. By Feature Reduction, the model complexity is reduced and it is easier to interpret. Moreover, the attenuation of the variables to collect is an advantage during the deployment of the model. In some cases, the variable selection enables to improve the model accuracy. Manual selection by an expert domain is certainly the best approach. But because the number of candidate descriptors is often large, it is not always possible in practice. [4]. So, we must select automatically the best variables. The various feature selection Algorithms that were tried includes:

i. Fisher filtering

It is a supervised feature selection Algorithms based upon a filtering approach i.e., processes the selection independently from the learning Algorithm. This component ranks the inputs attributes according to their relevance. This approach does not take into consideration the redundancy of the input attributes. [3]

ii. Relief filtering

This is a supervised Algorithm which will not consider the redundancy of the input attributes. At least two attributes must be available and the target attribute must be discrete. [3]

iii. Step Disc

Step disc is always associated to discriminant .We implement the FORWARD and the BACKWARD strategies in TANAGRA. In the FORWARD approach, at each step, we determine the variable that really contributes to the discrimination between the groups. We add this variable if its contribution is significant. In the BACKWARD approach, we begin with the complete model with all descriptors. We search which is the less relevant variable. We remove this variable if the removing does not significantly damage the discrimination between groups. The process stops when there is no variable to remove. [3]

IV. EXPERIMENTAL RESULTS

This section describes about the results obtained in the research work. The features selected by the above stated feature selection algorithms are displayed in the Table 2.

Table 2
Features selected under different algorithms

S.No	Feature selection algorithm	No.of features selected	Name of the features
1	Fisher filtering	6	Ascites, albumin, varices, histology, bilirubin, spiders
2	Relief filtering	9	Spiders, histology, ascites, fatigue, malaise, varices, live firm, steroid, spleen palpable
3	Step disc	4	Ascites, Spiders, Bilirubin, Anorexia

In this paper 15 classification algorithms shown in Table 3 are run in Tanagara tool using Hepatitis data set. Error rate is the performance measure for analyzing a data mining algorithm. In this paper, classifications algorithms are analyzed by applying and not applying the filters (filters in our case are the feature selection algorithms).

Table 3
Performance Analysis

Classification Alg	Without feature selection	With feature selection		
	Error Rate (All)	Fisher filtering	Relief Filtering	Step disc
BVM	0.0000	0.1071	0.0446	0.1250
C4.5	0.0893	0.1161	0.1250	0.1250
C-RT	0.1696	0.1696	0.1696	0.1696
CS-CRT	0.1696	0.1696	0.1696	0.1696
CS-MC4	0.1161	0.1161	0.1339	0.1339
C-SVC	0.0714	0.1339	0.1339	0.1339
CVM	0.000	0.0982	0.0446	0.1161
ID3	0.1696	0.1696	0.1696	0.1696
K-NN	0.1161	0.1518	0.1250	0.1071
LDA	0.0982	0.1429	0.1339	0.1250
MLP	0.0625	0.1161	0.1429	0.1250
NBC	0.1607	0.1696	0.1429	0.1518
PLS-DA	0.1071	0.1250	0.1339	0.1161
PLS-LDA	0.1071	0.1518	0.1339	0.1250
Rnd Tree	0.000	0.000	0.0625	0.0357

The performance of the algorithm with feature selection and without feature selection is shown in Figure 1 and Figure 2 respectively. From the above the analysis it is inferred that the classification algorithms works well without feature selection i.e. performance of the classification algorithm is high when all the attributes in a data set are present. The accuracy rate of the classification algorithms were not promising if the features selected by the feature selection algorithms as shown in table 2. Hence it is evident that the feature selection algorithm did not work well for this dataset and all the features are very essential for efficient classification of the dataset considered in this work.

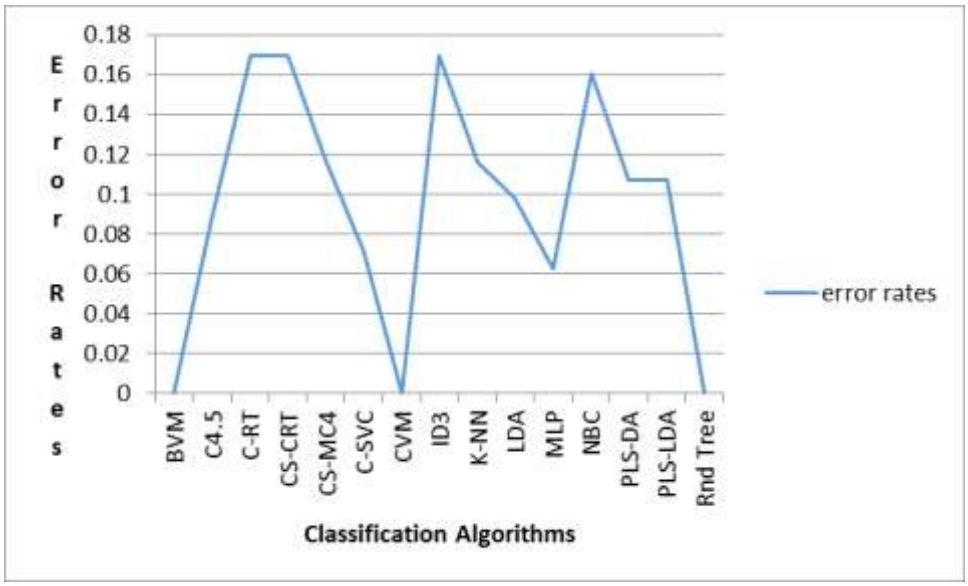

Figure 1. Performance of classification algorithms without feature selection

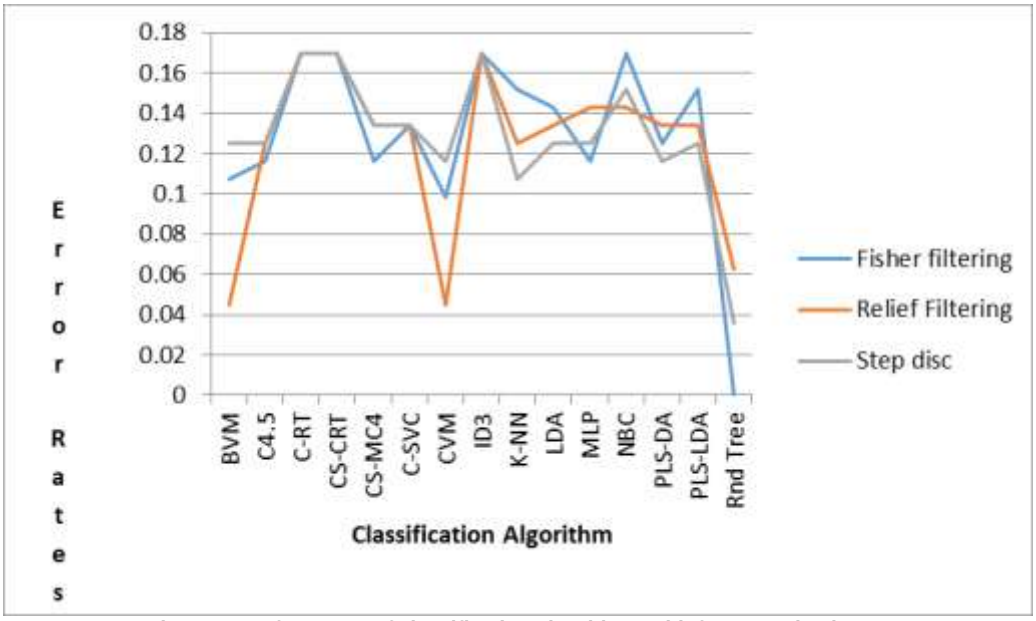

Figure 2. Performance of classification algorithms with feature selection

V. CONCLUSION

This research work mainly aims at highlighting the impact of data mining techniques on clinical data that may pave avenues for advancement in health-care and clinical decision-making. The comparison of classification techniques on predicting the survival of the patients was done using feature selection algorithms. We have analyzed the impact three feature selection algorithms towards classification efficiency and found that the algorithms towards feature selection did not yield promising results in classifying the hepatitis dataset. Analysis of many medical datasets will aid in discovering new patterns of disease occurrence and unearth the effect of data mining techniques in the medical arena.

REFERENCES

[1] International Journal of Computer Applications (0975 – 8887) Volume 62– No.15, January 2013 25 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.

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

[3] 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.

[4] Yılmaz Kaya, Murat Uyar, *A hybrid decision support system based on rough set and extreme learning machine for diagnosis of hepatitis disease.* 2013 Elsevier, 3429–3438

[5] Javad Salimi Sartakht, J. S. (2011). *Hepatitis disease diagnosis using a novel hybrid method*. Elsevier, 570-579.

[6] Duygu ʒCalisir, Esin Dogantekin, *A new intelligent hepatitis diagnosis system: PCA–LSSVM*, 2011 Elsevier, 10705–10708

[7] G.Sathyadevi, *Application of CART Algorithm in Hepatitis Disease Diagnosis*, 2011 IEEE, 1283-1287

[8] A.H.Roslina, & A.Noraziah. (2010). *Prediction of Hepatitis Prognosis Using Support Vector Machine and Wrapper Method*, IEEE, 2209-2211.

[9] Fadl Mutaher Ba-Alwi, H. M. (Volume 4, Issue 8, August-2013). *Comparative Study for Analysis the Prognostic in Hepatitis Data: Data Mining Approach*. International Journal of Scientific & Engineering Research, 680-685.