A Comprehensive Analysis on Preprocessing and Classification Techniques in Data Mining for Predicting Liver Disorder from USA Liver Patients Data

P.LauraJuliet¹, Dr.P.R.Tamilselvi²

¹Department of Computer Applications, Vellalar College for Women, Erode, (India)
²Department of Computer Science, Government Arts & Science College, Komarapalayam, (India)

ABSTRACT

Data mining is an activity of extracting some useful knowledge from a large data base, by using any of its techniques. Liver disorder is considered as one of the major problem which is unknowingly growing in modern society. The depth and severity of the disease will be rectified in matured stages, because the early symptoms are not much more effective. Data mining technique could be used for finding the type of Liver disorder using distinct symptoms. There are various data mining techniques and algorithms used for finding the liver disorder, namely Decision Tree, Naïve Bayes, Neural Network, Support Vector Machine, Multilayer Perceptron, Fuzzy Logic, and Hybrid Genetic Algorithm with various preprocessing techniques namely Data Cleaning, Data Integration, Data transformation and data reduction are analyzed on liver disorder data set. So the identification of liver disorder in initial phase is a vital activity to make awareness about the disease. These techniques are useful in timely and accurate classification and prediction of diseases and better care of patients. The main focus of this work is to analyse the use of data mining techniques by different authors for the prediction and classification of liver disorder. The dataset used in this work is Liver functional test data from USA liver patients.

Keywords—Data Mining, Preprocessing, Classification Techniques, liver disorder, USA Liver Patients Data.

I. INTRODUCTION

The liver is an important organ in our body. It weighs about 3 lb (1.36kg). It is reddish brown in colour and is divided into four lobes of unequal size and shape. Blood is carried to the liver via two large vessels called the hepatic artery and the portal vein.

Liver Disorder: Several diseases states can affect the liver. Some of the diseases are Wilson’s disease, hepatitis (an inflammation of the liver), liver cancer, and cirrhosis (a chronic inflammation that progresses ultimately to organ failure). Alcohol alters the metabolism of the liver, which can have overall detrimental effects if alcohol is taken over long period of time. Hemochromatosis can cause liver problems.
II. DATA MINING AND ITS TECHNIQUES

The term data mining refers to the findings of relevant and useful information from databases. This encompasses a number of technical approaches, such as clustering, data summarization, classification, finding dependency network, analysing changes and detecting anomalies. Data mining and knowledge discovery is the new interdisciplinary field, merging ideas from statistics, machine learning, database and parallel computing.

2.1 KDD

KDD is a process of extracting the hidden knowledge from data warehouse. Steps in, knowledge discovery data process are Selection, Pre-processing, data Transformation, data mining, data integration, Evaluation and Visualization. There are various data mining techniques like classification, clustering, association rule mining and prediction.

2.2 Classification Techniques in Data Mining

Classification is the process of finding a model that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data. Classification predicts categorical labels. Classification models are tested by comparing the predicted values to known target values. Classification is composed of two steps – training and testing.

C4.5[2][10]

C4.5 is an algorithm used to produce a decision tree which is increase of prior ID3 calculation. It enlarges the ID3 algorithm is managing both continuous and discrete properties, missing values and pruning trees after construction. The decision trees created by C4.5 can be used for grouping and often referred to as a statistical classifier. C4.5 creates decision trees from a set of training data same way as Id3 algorithm. As it is a supervised learning algorithm it requires a set of training examples which can be seen as a pair: input object and a desired output value (class). The algorithm analyzes the training set and frame a classifier that must have the size to accurately arrange both training and test cases[11].

ID 3

Id3 calculation starts with the authentic set as the root hub. On every cycle of the algorithm it indicate through every unused attribute of the set and figures the entropy (or data pick up IG(A)) of that attribute. At that point determine the attribute which has the smallest entropy (or biggest data gain) value. The set is S then split by the selected attribute (e.g. marks < 50, marks < 100, marks >= 100) to produce subsets of the information. The algorithm proceeds to recuse on each and every item in subset and considering only items never selected before [11].

CHAID[9]

It attempts to stop growing the tree before over fitting occurs, whereas CART, ID3 and C4.5 generate a fully grown tree and then carry out pruning as post processing step. In this sense CHAID avoids the pruning phase.

Random Forest[2]
Random Forests are an ensemble learning method also thought of as a form of nearest neighbor predictor for classification and regression that compose a number of decision trees at training time and outputting the class that is the mode of the classes output by single trees. Random Forests are a combination of tree predictors where each tree based on the values of a random vector sampled independently with the same distribution for all trees in the forest[11]

**Naive Bayes Classifier[1][2][3][8][10]**

The Naive Bayes Classifier technique is based on Bayesian theorem and is especially used when the dimensionality of the inputs is high. The Bayesian Classifier is adequate of calculating the most possible output based on the input. It is also possible to add new raw data at runtime and have improved probabilistic classifier. A naive Bayes classifier regard that the presence (or absence) of a particular feature (attribute) of a class is unrelated to the presence (or absence) of any other feature when the class variable is given [11].

**Support Vector Machine[4][8][10]**

SVM have attracted a great deal of attention in the last decade and actively tested to various domains applications. SVMs are mostly used for learning classification, regression or ranking function. SVM are based on statistical learning theory and structural risk minimization principal and have the intent of determining the location of decision boundaries also known as hyperplane that produce the optimal separation of classes. SVM is the most robust and exact classification technique, there are many problems. The data analysis in SVM is based on convex quadratic programming, and it is computationally costly, as solving quadratic programming methods require large matrix operations as well as time consuming numerical computations.[11]

**Multilayer Perceptron(MLP)[4]**

It is a development from the simple perceptron in which extra hidden layers are added. The presence of these layers allows an ANN to approximate a variety of non-linear functions. The transfer function generally a sigmoid function. Multilayer perceptron is a neural network that trains using back propagation.

**Bayesian Network**

A Bayesian network is a graphical model that encodes probabilistic relationships among variables of interest. When used in conjunction with statistical techniques, the graphical model has several choices for data modeling. One, because the model encodes dependencies among all variables, it freely handles situations where some data entries are missing. Two, a Bayesian network Classification can be used to learn causal relationships, and hence can be used to gain understanding about a problem domain and to predict the consequences of intervention. Three, because the model has both a causal and probabilistic semantics, it is an ideal representation for combining prior knowledge (which often comes in causal form) and data. Four, Bayesian statistical methods in conjunction with Bayesian networks offer an efficient and principled way for avoiding the over fitting of data[8].
III.LITERATURE REVIEW

In the year of 2010, P.Rajeswari,G.SophiaReena[1] "Analysis of Liver Disorder Using Data mining Algorithm” In this paper the data classification is based on liver disorder. The training dataset is developed by collecting data from UCI repository consists of 345 instances with 7 different attributes. The values are normalized. This paper deals with results in the field of data classification obtained with Naïve Bayes algorithms , FT tree algorithms, and KStar algorithms and on the whole performance made know FT Tree algorithm when tested on liver disease datasets , time taken to run the data for result is fast when compare to other algorithm with accuracy of 97.10%.

In the year of 2013, Sa’diyah Noor NovitaAlfisahrin, Teddy Mantoro[2] proposed "Data Mining Technique For Optimatization of Liver Disease Classification” This study aims to identify if the patients have the liver disease based on the 10 important attributes of liver disease using a Decision Tree, Naive Bayes , and NBTree algorithms. The result shows NBTree algorithm has the highest accuracy, however the Naïve Bayes algorithm gives the fastest computation time. For future study, the performance of NBTree algorithm will be the target of improvement of the accuracy by finding the most significant factor in identifying liver disease patients.

In the year of 2014, S.Dhamodharan proposed [3] "Liver Disease Prediction Using Bayesian Classification”. This paper they use Bayesian classification technique. There are many liver disorders that require clinical care of the physician. They predict three major liver diseases such as liver cancer , cirrhosis , hepatitis with the help of distinct symptoms. The primary goal is to predict the class types from classes such as liver cancer, cirrhosis, hepatitis and “no diseases”. In this paper Naïve Bayes and FT tree algorithm accuracy are compared and the result is obtained. The result concludes that the accuracy of Naïve Bayes algorithm is much better than the other algorithms with accuracy of 72.6624%.

In the year of 2015, A.S.AnneshKumar, Dr.C.JothiVenkateswaran proposed[5] “A Novel approach for Liver Disorder Classification using Data Mining Techniques” This paper describes the categorization of liver disorder through feature selection and fuzzy K-means classification. They preprocess the data set to avoid inconsistent and missing values. In any of the medical diagnosis activity, some features may directly or indirectly influence and some others may not influence. They used novel hybrid approach of feature selection chooses thirty physical and clinical attributes out of 48. They performed a classification accuracy analysis in these selected classes and obtained fuzzy based classification gives better performance with the accuracy 94%.
In the year of 2015, P.Thangarajul, R.Mehala [6] “Performance Analysis of PSO- K Star Classifier over Liver Diseases” The objective of this paper is to analyze the data of liver diseases using particle swarm optimization algorithm (PSO) with K Star classification. The proposed algorithm enhanced the performance of accuracy when compared to existing classification algorithms. PSO-Kstar algorithm is best suitable algorithm for the classification of liver disorders as it improved the performance in prediction accuracy is 100%.

In the year of 2015, Onwodi Gregory [7] proposed "Prediction Of Liver Disease (Biliary Cirrhosis) Using Data Mining Techniques". In this study, two real liver patient datasets were investigated for building classification models in order to predict liver diagnosis. Eleven data mining classification algorithms were applied to the datasets and the performance of all classifiers are compared against each other in terms of accuracy, precision, and recall. FT Tree algorithm when tested on liver disease datasets, time taken to run the data for result is fast when compared to other algorithms with 65% of accurate result. The RBF Network algorithm compared to other classification algorithms, it also shows the enhanced performance according to the attributes and it gives 78.0% of Accuracy, 77.5% of Precision, 86.4% of Sensitivity and 38.2% of Specificity results respectively.

In the year of 2015, Dr. S.Vijayarani, Mr. S.Dhayanand [8] proposed "Liver Disease Prediction using SVM and Naive Bayes Algorithms". The main objective of this research work is to predict liver diseases using classification algorithms. The algorithms used in this work are Naive Bayes and support vector machine (SVM). These classifier algorithms are compared based on the performance factors i.e. classification accuracy and execution time. From the experimental results it is observed that the SVM is a better classifier for predict the liver diseases. On the other hand, while comparing the execution time, the Naive Bayes classifier needs minimum execution time.

In the year of 2015, “Ebenezer Obaloluwa Olaniyi, khashman Aadnan”, [10] proposed “Liver Disease Diagnosis Based on Neural Networks”. In this paper back propagation neural network and radial basis function neural network are designed to diagnose these diseases and also prevent misdiagnosis of the liver disorder patients. In the preprocessing stage noisy data’s are cleaned using WEKA tool. The algorithms were compared with the c4.5, CART, Naive Bayes, Support Vector Machine (SVM) and concluded that the radial basis function neural network is the optimal model because it has a recognition rate of 70% which has proved more accurate and efficient than the other algorithms. In the year of 2016, Tapas Ranjan Baithara, Subhendu Kumar Pani proposed [9] “Analysis of Data Mining Techniques For Healthcare Decision Support System Using Liver Disorder Dataset”. This research focuses on the aspect of Medical diagnosis by learning pattern through the collected data of Liver disorder to develop intelligent medical decision support systems to help the physicians. Here normalization is done by dividing each sample of a particular attribute by its maximum value. In this paper the use of decision trees J48, Naive Bayes, ANN, ZeroR, 1BK and VFI algorithm to classify these diseases and compare the effectiveness, correction rate among them. By analyzing the results, Multilayer perceptron gives the overall best classification result with the accuracy 71.59% than other classifiers.
IV. CONCLUSION

Data mining is helpful for analysing the data, when the manually analysis of data is not feasible then the data mining techniques are applied for analysis. The data mining techniques are computer based algorithms which identify the relationship among the data and extraction of the similar pattern data on which they are trained. This paper presents a critical review of various data mining based techniques for the classification and prediction of liver disorder related data.

REFERENCES

JOURNAL PAPERS:


[10] Ebenezer Obaloluwa Olaniyikhashman Aadnan, “Liver Disease Diagnosis Based on Neural Networks”, Advances in Computational Intelligence, Proceedings of the 16th International Conference on Neural Networks (NN’15), November 7-9, 2015.


BOOKS:
