International Journal of Pure and Applied Mathematics
Volume 118 No. 9 2018, 765-770
ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue
ijpam.eu

Prediction and Analysis of Liver Disorder Diseases by using Data Mining Technique: Survey

Shambel Kefelegn
Research scholar, Computer Science and IT, Symbiosis Institute of Technology
Symbiosis International University
Pune, India
shamble.bereded@sitpune.edu.in

Pooja Kamat
Assistant professor (IT), Computer Science and IT,
Symbiosis Institute of Technology
Symbiosis International University
Pune, India
pooja.kamat@sitpune.edu.in

***Abstract*— Liver disorder diseases one of the major diseases in the world, Liver is one of the huge solid organ in the human body; and is also considered a gland because, among its many functions, it makes and secretes bile. The liver theatres vital role in many physical functions from protein manufacture and blood clotting to fat, sugar and iron metabolism. Liver disorder diseases are any trouble of liver purpose that reason for sickness. The experiment performed by using different data mining classification algorithm to get better result compared with the earlier liver diseases prediction study. All experiments are executed within the Weka tool. The study of paper to predicting and analysing liver disorder diseases to produce better performance accuracy by comparing various data mining classification algorithm and the performance of the accuracy is measured by confusion matrices.**

***Keywords*—Data Mining, SVM, NBC, C4.5, confusion matrices, liver disorder**

I. Introduction

One of the major cause of human death is liver disorder diseases. Liver is the second largest inside organ in the human body. Playing a key role in the metabolism and serving several imperative functions and its disorder has become one of the big issues of human diseases around the world. Liver diseases are one of the most killer diseases by the most cause of Viral Hepatitis in the world. Enhanced health analysis may be accomplished complete automatic diagnosis of patient record stored in health data, i.e., By learning from past experiences [1]. And data mining is applied to this stored medical record to retrieve information from the data [2]. According to [2] data mining techniques categorized into Association, classification, and clustering. Association its find the degree of the co-occurrence the item in the data, classification it's a data mining function that assigns the item to the target class, and it predicates the actual class to the items but Clustering unlike classification it's not assigned the class, and it divides the dataset into small partitions based on its similarity.

Data mining Techniques

As we mentioned before data mining techniques categorized into Association, classification, and clustering

A. Association

Association used to create an association between items and are often used to analyse sales transactions. The goals of association data mining to establishing a relationship between an item that occurs organized in a given dataset.

In data mining, association rules are used for analysing and guessing the medical health prediction to get a better diagnosis.

B. Clustering

The activity of clustering is dividing a set of record into a set of the meaningful homogeneous cluster. Clustering is the grouping organized of comparable records into clusters.the most popular data mining technique is clustering analysis; the technique of clustering algorithm will impact the clustering outcomes directly [23].

C. Classification

Classification is a model used to predict the future behaviour of the data through classifying the records into predefined classes. The classification algorithm is measured by in terms of precision and recall metrics to estimate the performance of classification algorithm. There are various data mining classifiers some of them are listed below: -

Naïve Bayes

Naive Bayes in the huge data set presented acceptable speed and accuracy, but the effect is outstandingly unfortunate in the instance of small dataset [4].and The NB classifier the probabilistic algorithm that calculates a set of probabilities by counting the frequency and groupings of values in a given records. [5]. The equation of Bayes Theorem states below:

$$Posterior = \frac{Prior * Likilihood}{Evidence}$$

$$P(C|X) = \frac{P(C) * P(C|X)}{P(X)}$$

Support vector machine

The Support Vector Machine (SVM) was first formed by Vapnik and has since involved a high grade of concentration in the machine learning study area [7]. Support Vector Machine is a constant algorithm compared to other algorithms that are neural networks, decision trees [8].

C4.5 Decision Tree

C4.5 is the most fundamental and the most well-organized classifier in decision tree-based method [6]. There are many types of data mining classification algorithm one of them is decision tree.C4.5 decision tree classifier is the upgraded form is the conventional decision tree classifier is IDE3 [9].

Algorithm	Advantage	Disadvantage
Naïve Bayes	Super simple Extremely fast High probability classification algorithm It predicts correct result Good performance	The NBC requires a very large number of records to obtain good results. It is instance-based or lazy in that they store all of the training samples
Support vector machine	Prediction result mostly better Fast estimation of the learned target Less parameter to consider	Computationally expensive, thus runs slow
C4.5 Decision tree	It produces the accurate result. It takes the less memory to large program execution. It takes less model build time. It has short searching time.	Empty branches. Insignificant branches. Overfitting.

Table 1 Comparison of Classification Algorithm

The paper is divided into five sections. Section 1 describes the introduction and algorithms associated with data mining. In Section 2, it describes the data mining challenges. In section 3, it describes literature review it compares the classification algorithms deeply. In Section 4, it provides the proposed methodology. Finally, in section 5 it describes the conclusion of the entire

II. LITERATURE REVIEW

Classification algorithm is one of the greatest significant and applicable data mining technique used to apply in disease prediction. Classification algorithm is the most common in several automatic medical health diagnoses. Many of them show a good classification accuracy listed below [10].

The Ahmed et al. [11] It predicts the breast cancer recurrence by using breast cancer dataset, and it compares using different classification algorithm like SVM, C4.5, and NBC. By comparing those algorithms, and by using SVM classification algorithm it got 75.75% compared to another classification algorithm it has high value. At this paper, it used K-fold technique for data partitioning.

The Hiba et al. [10] by using Machine learning algorithms for breast cancer and it uses breast cancer dataset, unlike other in this paper it compares the classification algorithms like KNN, C4.5, SVM and NB. By comparing those classifier SVM scores better accuracy 97.13% with lowest error rate.

The Roohallah et al. [12] a data mining approach for diagnosis of coronary artery diseases by using Z-Alizadeh Sani dataset and using data mining classification algorithm such as SVM, NBC, ANN, and Bagging Algorithm. The experiment shows that SVM is more accurate than other classification algorithm; it scores accuracy of 94.04%.

The Evandro et al. [13] Evaluating the effectiveness of students' academic failure in introductory programming courses this paper analysis student academic failure by using different classification algorithms such as NBC, J48, SVM, and ANN. By comparing those classifiers, NBC produces 87.12% accuracy with low error rate, and it uses tenfold cross-validation to partition data and confusion metrices to measure the performance accuracy. The Saba et al. [14] HMV: A medical decision support framework using multi-layer classifiers for disease prediction, this paper using HMV (hierarchical majority voting) ensemble framework for a medical decision analysis and the HMV ensemble framework utilizes F-score feature selection method. Accuracy, sensitivity, specificity and F-measure compression with another classifier HMV ensemble achieved the highest accuracy. It has achieved 71.53% for ILPD data set and 67.54% accuracy for Bupa liver diseases data set when compared with another classification algorithm.

The Veenita et al. [15] Chronic Kidney Disease Analysis Using Data Mining Classification Technique by using the original clinical dataset and it compares data mining classifiers such as NB and ANN. The experiment result shows NB classifier produce better accuracy 100% when compared with the Artificial neural network classification algorithm.

The U Rajendra et al. [16] Thyroid lesion classification in 242 patient population using Gabor transform features from high-resolution ultrasound images, this paper using Gabor transform feature and various types of classifier like SVM, K-NN, C4.5, and MLP. By comparing those classifier C4.5 decision tree produces 94.3% accuracy. The Prasad et al. [17] Implementation of Partitional Clustering to Predict Liver Disorders by using ILPD dataset and compares the performance of supervised learning classifiers; such as Naïve Bayes, C4.5 decision tree, and k- Nearest Neighbor; to find the best classifier in liver disorders diseases. The experiment results show that NB classifier produces highest accuracy 69%. The Randa et al. [18] Feature Analysis of Coronary Artery Heart Disease Data Sets, by using the heart diseases dataset and it uses C4.5 decision tree classifier. The c4.5 decision tree has produced 78.57% accuracy, and Ten-fold cross-validation used to the data partition.

The J. Pradeep et al. [20] Performance Analysis of Classifier Models to Predict Diabetes Mellitus, this paper predicts the diabetes mellitus by using the diabetes dataset and classification algorithms like J48, KNN, SVM and Random Forest, the performance of the classifiers has been measured by using two cases i.e., dataset with before pre-processing and dataset with after pre-processing such a Random forest classifier in both cases produce the highest accuracy compares with other classification algorithms. The M.khil et al. [21] Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm, this paper predicts heart diseases by using heart diseases dataset, and it uses KNN classification algorithm. The experiment shows that KNN gets better accuracy 100% with low error rate. The Tapas et al. [22] Analysis of Data Mining Techniques for Healthcare Decision Support System, by Using Liver Disorder Dataset and it compares the classification algorithms such as ANN, J48, NB, IBK, ZeroR, and VFI. The ANN classifier produces better accuracy 71.59% when compared with other classifiers.

Title	Author	Year	Method	Accuracy
Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses [13]	Evandro et al. [13]	2017	NBC, J48, SVM, ANN	87.12 %
HMV: A medical decision support framework using multi-layer classifiers for disease prediction [14]	Saba et al. [14]	2016	HMV ensemble	78.79 %
Chronic Kidney Disease Analysis Using Data Mining Classification Technique [15]	Veenita et al. [15]	2016	Naïve Bayes Artificial Neural Network	100%
Predicting Breast Cancer Recurrence using effective Classification and Feature Selection technique [11]	Ahmed et al. [11]	2016	SVM, NBC, C4.5	75.75 %
Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis, [10]	Hiba et al. [10]	2016	SVM, C4.5, NBC, K-NN	97.13 %
A data mining approach for diagnosis of coronary artery disease [12]	Roohallah et al. [12]	2013	SVM, NBC, ANN and Bagging Algorithm	94.08 %
Implementation of Partitional Clustering on ILPD Dataset to Predict Liver Disorders [17]	Prasad et al. [17]	2016	C4.5, NBC, K-NN	69%
Feature Analysis of Coronary Artery Heart Disease Data Sets Elsevier, International [18]	Randa et al. [18]	2015	C4.5	78.57 %
Performance Analysis of Classifier Models to Predict Diabetes Mellitus [20]	Pradeep et al. [20]	2015	K-NN, J48, SVM, Random forest	83.6%
Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm [21]	M.khil et al. [21]	2013	K-NN	100%

Table 2. Summary of Literature Review

With respect to all literature review cited above, associates the performance of data mining classification algorithms that is SVM, NB and C4.5 decision tree based on prediction and analysis to make a decision.

III. Proposed methodology

In the proposed methodology system, I decided to used three classification algorithms to predict the liver disorder diseases by comparing the performance accuracy of each classification algorithms. The classifiers are SVM, NB and C4.5 decision tree classifiers. K-fold cross-validation used to data partitioning based on Test set used to test the model and Training set used to train the data.

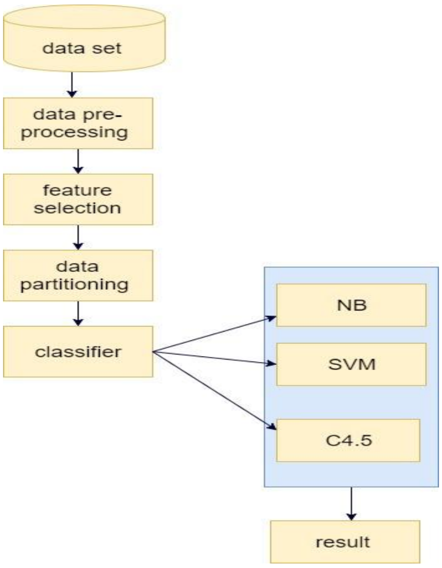

Fig1. Proposed Methodology Work Flow Diagram

IV. Conclusion

Data mining is the procedure of retrieve a pattern from large data set in connection with machine learning, data base, and statistics. A data mining technique such clustering, classification and association which is appropriate for medical diagnosis. Popular classification algorithm such as SVM, NB and C4.5 Decision Tree considered for performance evaluation in liver disorder diseases prediction. In liver disorder diseases there are 583 data sets with 10 attributes. The attributes are Total Bilirubin, Direct Bilirubin, Total Proteins, Albumin, A/G ratio, SGPT (Alamine Aminotransferase), SGOT (Aspartate Aminotransferase) and Alkaline Phosphatase.

Future work we can use the Hybrid approach to get better performance accuracy for liver disorder diseases prediction with their suitable data sets.

Reference

[1]. B. Tapas Ranjan, and Subhendu Kumar Pani. "Analysis of Data Mining Techniques for Healthcare Decision Support System Using Liver Disorder Dataset." *Procedia Computer Science* 85 (2016): 862-870.

[2]. P. Sheenal, and Hardik Patel. "Survey of data mining techniques used in healthcare domain." *International Journal of Information* 6, no. 1/2 (2016).

[3]. N. Eric WT, Li Xiu, and Dorothy CK Chau. "Application of data mining techniques in customer relationship management: A literature review and classification." Expert systems with applications 36, no. 2 (2009): 2592-2602.

[4]. H. Yuguang, and Lei Li. "Naive Bayes classification algorithm based on small sample set." In *Cloud Computing and Intelligence Systems (CCIS), 2011 IEEE International Conference on*, pp. 34-39. IEEE, 2011.

[5]. P. Tina R., and S. S. Sherekar. "Performance analysis of Naive Bayes and J48 classification algorithm for data classification." *International Journal of Computer Science and Applications* 6, no. 2 (2013): 256-261.

[6]. K. Masud, and Rashedur M. Rahman. "Decision tree and naive bayes algorithm for classification and generation of actionable knowledge for direct marketing." *Journal of Software Engineering and Applications* 6, no. 04 (2013): 196.

[7]. D. K. SRIVASTAVA, and B. Lekha. "Data classification using support vector machine." *Journal of Theoretical and Applied Information Technology* 12, no. 1 (2010): 1-7.

[8]. H. Ling-Min, Xiao-Bing Yang, and Hui-Juan Lu. "A Comparison of support vector machines ensemble for classification." In *Machine Learning and Cybernetics, 2007 International Conference on*, vol. 6, pp. 3613-3617. IEEE, 2007.

[9]. A. Rafik Khairul, and Yuliant Sibaroni. "Implementation of decision tree using C4. 5 algorithms in decision making of loan application by debtor (Case study: Bank pasar of Yogyakarta Special Region)." In *Information and Communication Technology (ICoICT), 2015 3rd International Conference on*, pp. 75-80. IEEE, 2015.

[10]. A. Hiba, Hajar Mousannif, Hassan Al Moatassime, and Thomas Noel. "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis." *Procedia Computer Science* 83 (2016): 1064-1069.

[11]. P. Ahmed Iqbal, Md Ahadur Rahman Munshi, Shahed Anzarus Sabab, and Shihabuzzaman Shihab. "Predicting breast cancer recurrence using effective classification and feature selection technique." In *Computer and Information Technology (ICCIT), 2016 19th International Conference on*, pp. 310-314. IEEE, 2016.

[12]. A. Roohallah, Jafar Habibi, Mohammad Javad Hosseini, Hoda Mashayekhi, Reihane Boghrati, Asma Ghandeharioun, Behdad Bahadorian, and Zahra Alizadeh Sani. "A data mining approach for diagnosis of coronary artery disease." *Computer methods and programs in biomedicine*111, no. 1 (2013): 52-61.

[13]. C. Evandro B., Baldoino Fonseca, Marcelo Almeida Santana, Fabrísia Ferreira de Araújo, and Joilson Rego. "Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses." *Computers in Human Behavior* 73 (2017): 247-256.

[14]. B. Saba, Usman Qamar, Farhan Hassan Khan, and Lubna Naseem. "HMV: a medical decision support framework using multi-layer classifiers for disease prediction." *Journal of Computational Science* 13 (2016): 10-25.

[15]. K. Veenita, Khushboo Chandel, A. Sai Sabitha, and Abhay Bansal. "Chronic Kidney Disease analysis using data mining classification techniques." In *Cloud System and Big Data Engineering (Confluence), 2016 6th International Conference*, pp. 300-305. IEEE, 2016.

[16]. A. U. Rajendra, Pradeep Chowriappa, Hamido Fujita, Shreya Bhat, Sumeet Dua, Joel EW Koh, L. W. J. Eugene, Pailin Kongmebhol, and Kwan-Hoong Ng. "Thyroid lesion classification in 242 patient population using Gabor transform features from high resolution ultrasound images." *Knowledge-Based Systems* 107 (2016): 235-245.

[17]. B. MS Prasad, M. Ramjee, Somesh Katta, and K. Swapna. "Implementation of partitional clustering on ILPD dataset to predict liver disorders." In *Software Engineering and Service Science (ICSESS), 2016 7th IEEE International Conference on*, pp. 1094-1097. IEEE, 2016.

[18]. E. Randa, Mostafa A. Salamay, Omar H. Karam, and M. Essam Khalifa. "Feature analysis of coronary artery heart disease data sets." *Procedia Computer Science* 65 (2015): 459-468.

[19]. K. J. Pradeep, and S. Balamurali. "Performance analysis of classifier models to predict diabetes mellitus." *Procedia Computer Science* 47 (2015): 45-51.

[20]. D. B. L., and Priti Chandra. "Classification of heart disease using k-nearest neighbor and genetic algorithm." *Procedia Technology* 10 (2013): 85-94.

[21]. B. Tapas Ranjan, and Subhendu Kumar Pani. "Analysis of Data Mining Techniques for Healthcare Decision Support System Using Liver Disorder Dataset." *Procedia Computer Science* 85 (2016): 862-870.

[22]. M. Madhiya. "An analysis on clustering algorithms in data mining." *International Journal of Computer Science and Mobile Computing* 3, no. 1 (2014): 334-340.