A Data Mining Approach to Prediction of Liver Diseases

Nazim Razali¹, Aida Mustapha¹, Mohd Helmy Abd Wahab², Salama A Mostafá¹, Siti Khadijah Rostam¹

¹ Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Parit Raja 86400 Batu Pahat, Johor, Malaysia
² Faculty Electrical and Electronic Engineering, Universiti Tun Hussein Onn Malaysia, Parit Raja 86400 Batu Pahat, Johor, Malaysia
E-mail: aidam@uthm.edu.my

Abstract. One of the major diseases in the world is liver disease. The liver is an important part of the large organs in the human body and is also considered a gland. This is because it creates and bites bile. Liver disease is a liver problem that causes the disease. The objective of this study is to propose a rule-based classification model with basic decision making techniques to predict various types of heart disease. To get better results the experiment was done using a different data mining algorithm compared with previous liver disease predictions. All experiments have been implemented at Azure Machine Learning tool. This paper is about to study the prediction of liver disease to produce better performance accuracy by comparing various mining data classification algorithms.

1. Introduction
Liver is the second largest inside organ in the human body. It plays an important role in human body such as for producing protein, clotting blood, as well metabolizing cholesterol, glucose, and iron. Liver also has the function for removing toxins from the body, hence is significant to ensure survival. When a liver fails to operate, many of the body functions cannot be well performed, hence causing significant damage to the body. Liver can be damaged if it is infected with a virus, attacked by its own immune system or injured by chemicals. One fatal liver disease is caused by hepatotrophic virus such as the hepatitis B virus (HBV), hepatitis C virus, and hepatitis delta virus, which can result in chronic liver disease.

Liver disease is also referred to as hepatic disease. Hepatic diseases produce symptoms such as nausea, vomiting, fatigue, abdominal pain and swelling, back pain, fatigue and weight loss. Certain patients are reported to suffer from jaundice (yellowing of the skin and eyes), fluid in abnormal cavity, pale stool, and especially enlarged spleen as well as gallbladder. Tests such as imaging tests and liver function tests can check for liver damage and help to diagnose liver diseases whether it is acute or chronic [1]. The definition of acute liver disease based on duration, with the history of the disease does not exceed six months.

According to [2], liver disease is one of the killer diseases in the world. To contain the disease requires an enhanced health analysis through automatic diagnosis of patient record stored in health institutions or organizations. A data mining approach can be used to classify the liver
disease into acute or chronic based on the patients’ symptoms. This allows the doctors or medical providers to extract the correct information in order to suggest for effective medical assistance.

The remaining of this paper organized as follows. Section 2 reviews all works related to the liver disease. Section 3 presents the methodology used to perform the data-mining task along with the dataset and the evaluation metrics. Section 4 presents the results and finally Section 5 concludes with some direction for future work.

2. Related Work

Data mining techniques can be categorized into type of analysis being carried out such as association rule mining, classification or clustering. Association rule mining looks for the degree of co-occurrence the symptoms in the health records and predicts the medical health prediction in order to get better diagnosis. Classification assigns the item to the target class, and it predicated the actual class to the items. Classification model is often used to predict the future behavior of the data and classify the data into class. Various data mining algorithms used such as Naïve Bayes, Support Vector Machines, and C4.5 Decision Tree. The algorithms used is to predict the liver disorder disease by comparing the accuracy of each algorithm [4]. Finally, clustering divides or clusters the patient symptoms into small partitions based on its similarity. It groups the organized of comparable records into clusters and the clustering analysis will affect the clustering outcomes directly.

[3] explored three different classification algorithms in classification, which are Naïve Bayes, KStar, and FT Tree. There were 7 attributes for classification. The experiments revealed that FT Tree algorithm has better performance because it takes only some time to complete the process and calculate the accuracy than Naïve Bayes and KStar. More data mining algorithms were explored in [5] such as the Decision Tree J48, Naïve Bayes, Multilayer Perceptron, ZeroR, K-Nearest Neighbor, and VFI algorithms to classify the liver disease. Based on the result analysis, the highest of the accuracy among those algorithms is Multilayer Perceptron (71.59%). It gives overall best classification result than other algorithms.

Comparative analysis on clustering and classification algorithms was also carried out by [6]. The result analysis showed that the classification is better than clustering algorithms with an accuracy of about 81%. In another work, classification and regression experiments were carried out based on 11 liver disease symptoms as features. Logistic regression, Bayes point machines, and two-class neural network algorithms used for classification approach while Linear Regression and Poisson Regression used for Regression approach for liver disease detection. The result analysis with respect to accuracy Logistic Regression algorithm has the highest accuracy whereas with respect to computational time Bayes Point Machines algorithm performs better.

Other than data mining the liver disease, [7] used data mining approach for diagnosis of coronary artery disease. Various classification algorithms were used such as the Support Vector Machine, Naïve Bayes Classification, Artificial Neural Network, and Bagging Algorithm. Based on the experiment that was carried out, the results showed that Support Vector Machines was more accurate that other algorithms with 94.04% of accuracy. [8], however, reported that Naive Bayes performed betther in the Artificial Neural Network (ANN).

K-Nearest neigbour (KNN) is another simple yet highly efficient data mining algorithm especially for classification and pattern recognition. [9] used a hybrid algorithm that combined Genetic Algorithm (GA) and KNN algorithm to improve the accuracy of classification of heart disease dataset. GA belongs to the class of evolutionary computing that has powerful search to reduce redundant and irrelevant attributes or features in a dataset, hence optimizing the results. The experiments also showed that by integrating GA with KNN, the experiment produced greatest accuracy rather than just the KNN algorithm.

More hybrid data mining methods was explored by [10] such as classification with clustering technique. For example, a K-Nearest Neighbors algorithm was integrated with Fuzzy C-Means
clustering. It is clear that the Fuzzy K-Nearest Neighbors with Fuzzy C-Means model produced the better result than the K-Nearest Neighbors with Fuzzy C-Means model on liver disorder datasets. The paper also concluded that the use of Fuzzy C-Means clustering algorithm for pre-processing improved the result in terms of classification accuracy along with better speed as the algorithm works by reducing the number of features from the original datasets. From experiment, it is been found that Fuzzy K-Nearest Neighbors with Fuzzy C-Means have accuracy of 96.13% and K-Nearest Neighbors with Fuzzy C-Means with accuracy of 98.95.

3. Methodology

Data mining used to discover pattern of diseases and makes effective decisions with the help of different machine learning. Data mining is a process of discovering data that stored electronically and automatically by computer. Data mining is about solving problems by analyzing data already present in databases. Data preprocessing is one of the most critical steps in data mining process. The sequences of steps identified in extracting knowledge from data are data cleaning, data integration, data transformation, and data reduction [11].

This study will use classification and regression as data mining tasks. Classification is used for predicting responses that can have just a few known values based on column in the dataset. Regression can predict one or more continuous variables. In classification, models are evaluated using Bayes point machine and neural network as algorithms, while regression using linear regression and Poisson regression. The classification methodology is shown in Figure 1.

The experiments were carried out using the Azure Machine Learning tool (https://studio.azureml.net) with 10-fold validation method for training and testing. Cross validation randomly divides the training data into a number of folds. When the building and evaluation process is complete for all folds, cross-validate model generates a set of performance.

3.1. Dataset

This study used dataset from the University of California Irvine (UCI) repository. This data sets contains 583 people which 416 liver patients records and 167 non liver patient records. The excerpt of the dataset is shown in Figure 1.

![Figure 1. Liver disease dataset](image)

3.2. Algorithms

The classification of liver disease in this paper will be performed using two regression algorithms, which are Linear Regression and Poisson Regression as well as two classification algorithms,
which are Neural Networks and Bayes Point Machines.

• Linear Regression Algorithm. The formula is shown in Equation 1.

\[
R^2 = 1 - \frac{SSres}{SStot} = SSot - \frac{SSres}{SStot} = \frac{SSreg}{SStot}
\]  

\[1\]

• Poisson Regression Algorithm. The formula is shown in Equation 2.

\[
Pr(Y_i = y_i | \mu_i, t_i) = \frac{e^{-\mu_i(ut)}y}{y!}
\]  

\[2\]

• Bayes Point Machine. The formula is shown in Equation 3.

\[
W = \sum_{i=1}^{m} ax
\]  

\[3\]

• Neural Network. The formula is shown in Equation 4.

\[
J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} \sum_{k=1}^{k} y(i) \log(h(\theta x)) + \frac{\lambda}{2m} \sum_{i=1}^{L-1} \sum_{j=1}^{j+1} (\theta_j)^2
\]  

\[4\]

3.3. Evaluation Metrics

The evaluation metrics for regression used in the experiments are mean absolute error, root mean squared error, relative absolute error, and coefficient of determination. For classification, the evaluation metrics are accuracy, precision, recall and F-Measure.

• Mean absolute error measures the average magnitude of the errors in a set of predictions. The formula for calculating mean absolute error measures is shown in Equation 5.

\[
MAE = \frac{1}{n} \sum_{j=1}^{n} |y_j - \hat{y}_j|
\]  

\[5\]

• Root mean squared error quadratic scoring rule that also measures the average magnitude of the error. The formula for calculating root mean squared error is shown in Equation 6.

\[
RMSE = \sqrt{\frac{1}{n} \sum_{j=1}^{n} (y_j - \hat{y}_j)^2}
\]  

\[6\]

• Relative absolute error is the average of the actual values. The formula for calculating relative absolute error is shown in Equation 7.

\[
RAE = \frac{\sum_{i=1}^{n} |p_i - a_i|}{\sum_{i=1}^{n} |a - a_i|}
\]  

\[7\]

• Coefficient of determination is square of the correlation between predicted scores and actual scores, it ranges from 0 to 1. The formula for calculating coefficient of determination is shown in Equation 8.

\[
R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}
\]  

\[8\]

• Accuracy is a ratio number of correctly classified instances to the total number of instance. The formula for calculating accuracy is shown in Equation 9 where, TP = True Positive, FP = False Positive, FN = False Negative, TN = True Negative.
\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}
\]

- Precision is ratio of actual true predicted instance out of total true instance. The formula for calculating precision is shown in Equation 10.

\[
\text{Precision} = \frac{TP}{TP + FP} \tag{10}
\]

- Recall is a ratio of actual true instance out of all items which are true. The formula for calculating recall is shown in Equation 11.

\[
\text{Recall} = \frac{TP}{TP + FN} \tag{11}
\]

- F-measure is a harmonic mean of both precision and recall. The formula for calculating F-measure is shown in Equation 12.

\[
F - \text{Measure} = \frac{2(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \tag{12}
\]

4. Results and Discussion
The purpose of the experiments is to compare the performance of linear regression and poisson regression algorithms in Liver Disease dataset. The results are shown in Table 1. The results showed that poisson regression have lower score than linear regression. It measures how close the predictions are to actual outcomes. The value coefficient of determination of poisson regression shows that the accuracy is more accurate than linear regression as shown in Figure 2.

| Algorithm          | Mean Absolute Error | Mean Squared Error | Relative Absolute Error | Relative Squared Error | Coefficient of Determination |
|--------------------|---------------------|-------------------|-------------------------|------------------------|-----------------------------|
| Linear Regression  | 0.382265            | 0.432015          | 0.985699                | 0.962514               | 0.037486                    |
| Poisson Regression | 0.375341            | 0.423920          | 0.967844                | 0.92678                | 0.073220                    |

Figure 2. Comparison of results
Next, the comparative experiments were repeated to compare the classification accuracy of two data mining algorithms, which are the Neural Network and Bayes Point Machines algorithm using the same Liver Disease dataset. The selected classifier run in different scenarios of the dataset. By analyzing the results, for classification algorithm, the overall best result is Bayes Point Machines. The results shown in Table 2. The results showed that Bayes Point Machines algorithm is the best algorithm used for solve the problem relating to the liver disease. It is because the algorithm has the best accuracy than other algorithm (refer to Figure 3).

| Algorithms                  | Accuracy | Precision | Recall | F-Score |
|-----------------------------|----------|-----------|--------|---------|
| Neural Networks             | 0.6685   | 0.3903    | 0.6087 | 0.4735  |
| Bayes Point Machines        | 0.7052   | 0.1000    | 0.0301 | 0.0550  |

Figure 3. Comparison results between algorithms

Figure 4 shows the graph of the result that shows the number of patient with and without liver disease disorder. 416 patients are having liver disease while 167 are not having the disease.

Figure 4. Number of patient with and without liver disease
5. Conclusions

Data mining is one of the processes of sorting large data sets to identify patterns that involve methods of machine learning, statistics, and database systems. Data mining techniques such as clustering, classification and association mining rule are not only apply in medical diagnosis, but also in the other field. Classification algorithms such as Support Vector Machine, Naive Bayes and Decision Tree are most used to consider the performance evaluation in liver diseases prediction. In liver diseases, there are 583 datasets which 416 records of liver patients and 167 non-liver patient records with 11 attributes. The attributes are age, gender, Total Bilirubin, Direct Bilirubin, Alkaline Phosphatase, Alanine Aminotransferase, Aspartate Aminotransferase, Total Proteins, Albumin, A/G ratio, an expert. Hybrid approach can be applied in future work in order to get better accuracy of results for liver diseases prediction.

Acknowledgements

This research is funded by Universiti Malaysia Perlis.

References

[1] Nahar N and Ara F 2018 International Journal of Data Mining and Knowledge Management Process 8
[2] Vijayarani S and Sudha S 2013 International Journal of Computer Applications and Information Technology 2 17–21
[3] Rajeswari P and Reena G S 2010 Global journal of computer science and technology
[4] Shamble K and Pooja K 2018 International Journal of Pure and Applied Mathematics 118 765–770
[5] Baitharu T R and Pani S K 2016 Procedia Computer Science 85 862–870
[6] Ojha U and Goel S 2017 International Conference on Cloud Computing, Data Science and Engineering, IEEE, 2017
[7] Alizadehsani R, Habibi J, Hosseini M J, Mashayekhi H, Boghrati R, Ghandeharioun A and Sani Z A 2013 Computer methods and programs in biomedicine 111 52–61
[8] Kunwar V, Chandel K, Sabitha A S and Bansal A 2016 Cloud System and Big Data Engineering (Confluence), 2016 6th International Conference, IEEE 300–305
[9] Deekshatulu B L and Chandra P 2013 Procedia Technology 10 85–94
[10] Chetty N, Vaisla K S and Patil N 2015 International Conference on Advances in Computing and Communication Engineering, IEEE 568–572
[11] Han J, Kamber M and Pei J 2015 Morgan Kaufmann Series in Data Management Systems 230–240