Informatics in Medicine Unlocked 9 (2017) 93–106

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu

Computer based prognosis model with dimensionality reduction and validation of attributes for prolonged survival prediction

C.G. Raji [a,*], H.S. Anand [b], S.S. Vinod Chandra [c]

[a] Department of Computer Science & IT, MEA Engineering College, Kerala, India
[b] Department of Computer Science, Muthoot Institute of Technology and Science, India
[c] Computer Centre, University of Kerala, India

ARTICLE INFO

Keywords:
Liver transplantation
Data preprocessing
Data mining
Principal component analysis
Ranking
Association rule mining algorithms
Survival prediction model

ABSTRACT

Medical databases contain large volume of data about patients and their clinical information. For extracting the features and their relationships from a huge database, various data mining techniques need to be employed. As Liver transplantation is the curative surgical procedure for the patients suffering from end stage liver disease, predicting the survival rate after Liver transplantation has a big impact. Appropriate selection of attributes and methods are necessary for the survival prediction. Liver transplantation data with 256 attributes were collected from 389 attributes of the United Nations Organ Sharing registry for the survival prediction. Initially 59 attributes were filtered manually, and then Principal Component Analysis (PCA) was applied for reducing the dimensionality of the data. After performing PCA, 197 attributes were obtained and they were ranked into 27 strong/relevant attributes. Using association rule mining techniques, the association between the selected attributes was identified and verified. Comparison of rules generated by various association rules mining algorithm before and after PCA was also carried out for affirming the results. The various rule mining algorithms used were Apriori, Treap mining and Tertius algorithms. Among these algorithms, Treap mining algorithm generated the rules with high accuracy. A Multilayer Perceptron model was built for predicting the long term survival of patients after Liver transplantation which produced high accuracy prediction result. The model performance was compared with Radial Basis Function model to prove the accuracy of survival of liver patients'. The top ranked attributes obtained from rule mining were fed to the models for effective training. This ensures that Treap mining generated associations of high impact attributes which in-turn made the survival prediction flawless.

1. Introduction

Liver transplantation is the ultimate curative therapy for people suffering from end stage liver disease. With the various advancements in the field of Liver transplantation, the survival rate is increasing day by day and patients depend more on this surgical treatment [1]. For Liver transplantation, getting appropriate donor-recipient match is very difficult. Predicting the survival of patients after Liver transplantation depends upon appropriate donor-recipient matching. Donors as well as recipients have a set of features and these features should be matched with each other. Manual method for doing donor-recipient features matching is a tedious task. In many research problems, the dimensionality of input is excessively large. This can make the process of mining relatively slow. One solution for this is to remove the attributes which are less significant. To find relevant attributes from the huge dataset, various data mining techniques can be employed.

One of the key areas of rule mining is association rule learning which helps to find the relationship between various attributes in a dataset. Association rules provide the co-occurrence of data items as well as the directionality between data items [2]. The association mining result can be used in creating a prediction model for long term survival. The data collected from UNOS consist of 389 clinical and non-clinical attributes. From the whole dataset, Liver transplantation attributes of 256 donor, recipient and transplantation details were extracted. But all of these attributes don't have a role to play in computational prediction. Selecting the relevant attributes of donor-recipient manually for the survival prediction in liver transplantation is not easy. But from any dataset, it is possible to manually filter some attributes of zero help even without any data mining knowledge or techniques. A total of 59 such attributes comprising donor, recipient and transplantation attributes were

* Corresponding author.
E-mail addresses: rajicg80@gmail.com (C.G. Raji), anandhareendrans@mgits.ac.in (H.S. Anand), vinod@keralauniversity.ac.in (S.S. Vinod Chandra).

http://dx.doi.org/10.1016/j.imu.2017.07.002
Received 2 November 2016; Received in revised form 5 July 2017; Accepted 6 July 2017
Available online 11 July 2017

manually eliminated.

Data mining techniques and association rule mining algorithms play a significant role in mining appropriate donor, recipient and transplantation attributes from the remaining 197 attributes. PCA was carried out for 197 attributes for dimensionality reduction. By ranking the resultant attributes with standard deviation, 27 attributes out of 197 attributes in the United Nations Organ Sharing (UNOS) Liver transplantation dataset were considered as relevant. To mine the relationship between attributes, various association rule mining algorithms like Apriori, Tertius and Treap mining algorithms were used. Results obtained were used as the training set for creating the Multilayer Perceptron (MLP) and Radial Basis Function (RBF) prediction models for survival.

Since the technological advancements resulted in the improved survival rate, liver patients depend upon Liver transplantation in-spite of the costly surgery. To integrate the expected post-transplant prediction into decisions about organ allocation, and to enable knowledgeable decision-making by potential liver patients and their relatives [3], it is required to precisely evaluate the possibility of post-transplant survival based on information that is available before transplantation [3]. Even though there were some efforts earlier to create a model for prediction, it couldn't meet the needed accuracy [3]. The main reason for this was the unsuitable selection of a prediction model. Survival prognosis is a complex nonlinear relationship affected by many interactive factors, especially for a complicated organ transplantation procedure. Artificial Neural Network (ANN)s achieves a better part by escaping local optimum over conventional statistical techniques and logistic regression models. Researchers have surveyed the survival rates at various periods; none of them evaluated the accuracy of models for a long term survival of liver transplantation. With the properly mined and ranked set of attributes of liver transplantation (27 attributes), survival prediction of liver patients in post transplantation was conducted. We proposed two ANN models such as MLP and RBF models for comparing the accuracy of survival of patients after Liver Transplantation. A successful survival prediction model with high accuracy was obtained as the self-learning capability of ANN (Artificial Neural Network) was exploited for model building. For successful survival prediction, follow up information after Liver transplantation is necessary.

To summarize, objectives of the study are:

1. To reduce the dimensionality of data set attributes and validation of attributes.
2. To find out significant and essential attributes by ranking.
3. Prove the relevancy of ranked attributes by creating a prediction model.
4. Perform the comparison of model to prove the survival accuracy.

2. Related research

With the development of health informatics, the amount of data available is increasing day by day. For finding the relevant data from a set of features, researchers introduced the concept of association rule mining algorithms. Adam et al. introduced an automated technique for identifying associations among laboratory results, medications and problems. The authors collected data from Brigham and Women's Hospital, Boston, MA in 2009 [4]. They generated associations using support, confidence, interest, chi square and conviction statistics. The study was single sited and the authors failed to find out infrequent occurring laboratory results, medications and problems in the dataset [4]. In 2012, Tzu-Chuen Lu et al. proposed a data mining technique which is used to determine the association rules from each cluster of hemo-dialysis patient data [5]. They collected the data from a hospital in Taiwan [5]. But they couldn't provide their complete dataset and it was very difficult to generate association rules from dataset. To detect and diagnose kidney disease and heart problems, Anu Choudhary et al. developed a prediction system using data mining techniques in 2014 [6]. They used 42 attributes with k-means clustering algorithm and Apriori algorithm [6]. They found that k-means clustering algorithm is more efficient than Apriori algorithm [6]. But they failed to extend the study to other diseases and transplantations. Mudathir et al. developed a program based on Apriori algorithm for extracting association rules from the renal transplantation data [7]. The data were collected from Cardiac Surgery & Renal Transplantations Center at Ahmed Qassim hospital in Khartoum North and the association tool used was Clementine and ARMADA [7]. They also conducted a study about renal failure in cardiovascular surgery patients in 2014 [7].

3. Materials and methods

3.1. Dataset

The dataset used to perform the survival prediction after Liver transplantation was collected from United Nations Organ Sharing (UNOS) registry [8]. It is the only administer of Organ Procurement & Transplant Network (OPTN), which is an official U.S Government network run by Health Resources and Services Administration (HRSA), U.S Department of health & Human Services. It is authenticated for medical dataset which is highly useful in prediction purposes. The UNOS database consists of 65,535 liver patient records with 389 attributes from 1st October 1987 to 5th June 2015. From the 389 attributes of UNOS database, 256 attributes of Liver transplantation was selected. The 256 attributes consist of recipient attributes, donor attributes and transplantation attributes. The dataset consists of both clinical and non-clinical details of male and female liver patients [9]. The dataset also consists of adult liver patient records and pediatric liver patient records. The waiting list information, Registration information, Transplantation information, Follow up information after Liver transplantation, Malignancies reported after Liver transplantation and recipient immune suppression at discharge are available in separate files. The records of recipient and donor are also available in separate files. All the records are linked using patient ID code and donor ID code. The liver transplantation dataset includes 5971 re-transplantation records also. The liver transplantation master file consists of all the recipient, donor and transplantation records with 256 attributes. The clear information about the Liver transplantation dataset was made available with the help of list of attributes, data type of attributes and description of attributes. Many of the attributes in the UNOS dataset are not significant in computational survival prediction of Liver transplantation [10]. In order to find attributes that contribute Liver transplantation survival prediction, association rule mining algorithms were used.

3.2. Data preprocessing

Not all attributes in the UNOS database consisting 389 attributes are useful for computational prediction of survival after Liver transplantation. Manually 59 attributes from the 256 attributes were removed during preprocessing. The 59 attributes consist of both recipient and donor attributes. Manually extracted attributes consist of 51 recipient attributes and 8 donors' attributes which are given in Table 1 of Supplementary File 1. All the transplantation attributes were having some relevance to the further processing and hence they were not eliminated. Some of the demographic features of recipient and donor, region, income and mode of payment also have no relevance in making a prediction. Thus such attributes were also manually removed.

3.3. Principal component analysis (PCA) and ranking the attributes

It is a non-parametric mathematical technique for mining relevant data from large datasets. PCA is applied to high dimensional data. PCA includes getting the covariance matrix, getting the eigen value and eigen vectors of each attributes and selection of main attributes [11]. With eigen analysis, the eigen values and eigen vectors of a square symmetric matrix with sum of squares and cross products can be calculated. The first attribute extracted in a principal component analysis accounts for a