Research Journal of Applied Sciences 10 (9): 469-473, 2015
ISSN: 1815-932X

Data Mining Association Rules for Heart Disease Prediction System

[1]R. Thanigaivel and [2]K. Ramesh Kumar
[1]Department of Computer Science and Engineering, [2]Department of Information Technology,
Hindustan University, Chennai, Tamilnadu, India

Abstract: Data mining techniques have been applied magnificently in many fields including business, science, the web, cheminformatics, bioinformatics and on different types of data such as textual, visual, spatial, real-time and sensor data. Medical data is still information rich but knowledge poor. There is a lack of effective analysis tools to discover the hidden relationships and trends in medical data obtained from clinical records. This study reviews the state of the art research on heart disease diagnosis and prediction.

Key words: Data mining, heart disease, prediction, web, diagnosis

INTRODUCTION

Heart and blood vessel diseases called cardiovascular diseases include numerous problems many of which are related to a process called atherosclerosis. Atherosclerosis is a condition that develops when a substance called plaque builds up in the walls of the arteries. This build-up narrows the arteries making it harder for blood to flow through. This can cause blood clot formation which can cause a heart attack (also called Myocardial Infarction) or stroke. When a heart attack occurs, the speed of detection and quick intervention is highly essential to save the life of heart attack patient and to prevent heart damage. Nowadays, the use of computer technology in the field of medicine has highly increased (Koutsojannis and Hatzilygeroudis, 2007). The use of intelligent systems such as neural network, fuzzy logic, genetic algorithm and neuro-fuzzy systems has highly helped in complex and uncertain medical tasks such as diagnosis of diseases (Patel *et al*., 2013). Over the last few decades, neural networks and fuzzy systems have established their reputation as alternative approaches to intelligent information processing systems. Both have certain advantages over classical methods, especially when vague data or prior knowledge is involved. However, their applicability suffered from several weaknesses of the individual models. Therefore, combinations of neural networks with fuzzy systems have been proposed where both models complement each other.

Neuro-fuzzy hybridization results in a hybrid intelligent system that synergizes these two techniques by combining the human-like reasoning style of fuzzy systems with the learning and connectionist structure of neural networks (Lin *et al*., 2009). The basic idea of combining fuzzy systems and neural networks is to design an architecture that uses a fuzzy system to represent knowledge in an interpretable manner and the learning ability of a neural network to optimize its parameters (Soni *et al*., 2011).

In year 2013, Vijiyaraniet performed a work, "An efficient classification tree technique for heart disease prediction". The data mining can be defined as discovery of relationships in large databases automatically and in some cases it is used for predicting relationships based on the results discovered. Data mining plays a vital role invarious applications such as business organizations, e-Commerce, health care industry, scientific and engineering. In the health care industry, the data mining is mainly used for predicting the diseases from the datasets. Various data mining techniques are available for predicting diseases namely classification, clustering, association rules and regressions. This study analyzes the classification tree techniques in datamining. The aim of this study is to investigate the experimental results of the performance of different classification techniques for a heart disease dataset. The classification tree algorithms used and tested in this research are decision stump, random forest and LMT Tree algorithm. Comparative analysis is done by using Waikato Environment for Knowledge Analysis or in short, WEKA. It is open source software which consists of a collection of machine learning algorithms for data mining tasks.

In year 2012, Koutsojannis and Hatzilygeroudis (2007) performed a work, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques". The health care industry is generally "information rich" but unfortunately not all the

Corresponding Author: R. Thanigaivel, Department of Computer Science and Engineering, Hindustan University, Chennai, Tamilnadu, India

data are mined which is required for discovering hidden patterns and effective decision making. Advanced data mining techniques are used to discover knowledge in database and for medical research, particularly in heart disease prediction. This study has analyzed prediction systems for heart disease using more number of input attributes. The system uses medical terms such as sex, blood pressure, cholesterol like 13 attributes to predict the likelihood of patient getting a heart disease. Until now, 13 attributes areused for prediction. This research paper added two moreattributes, i.e., obesity and smoking. The data mining classification techniques, namely Decision Trees, Naive Bayes and Neural Networks are analyzed on heart disease database. The performance of these techniques is compared, based on accuracy. As per our results accuracy of Neural Networks, Decision Trees and Naive Bayes are 100, 99.62 and 90.74%, respectively. Our analysis shows that out of these three classification models Neural Networks predicts heart disease with highest accuracy.

In year 2013, Vijiyaraniet performed a work, "An efficient classification tree technique for heart disease prediction". The data mining can be defined as discovery of relationships in large databases automatically and in some cases it is used for predicting relationships based on the results discovered. Data mining plays a vital role in various applications such as business organizations, e-Commerce, health care industry, scientific and engineering. In the health care industry, the data mining is mainly used for predicting the diseases from the datasets. Various data mining techniques are available for predicting diseases namely classification, clustering, association rules and regressions. This study analyzes the classification tree techniques in datamining. The aim of this study is to investigate the experimental results of the performance of different classification techniques for a heart disease dataset. The classification tree algorithms used and tested in this research are Decision Stump, Random Forest and LMT Tree algorithm. Comparative analysis is done by using Waika to Environment for Knowledge Analysis or in short, WEKA. It is open source software which consists of a collection of machine learning algorithms for data mining tasks.

The mining association rule usually adopts this model (Lin *et al*., 2009): support, confidence, interestingness. But this model can't measure the correlative degree between the antecedent and the consequent of the rule by ration. So, they proposed a new mining model of association rules: support, coincidence, interestingness and analyzed the meaning of coincidence by instance. At last, they used this model in the data about coronary heart disease and obtained a lot of meaningful rules.

The healthcare environment is still "information rich but" knowledge poor (Soni *et al*., 2011). There is a wealth of data available within the healthcare systems. However, there is a lack of effective analysis tools to discover hidden relationships and trends in data. This research intends to provide a of current techniques of knowledge discovery in databases using data mining techniques that are in use in today's medical research particularly in heart disease prediction. Number of experiment has been conducted to compare the performance of predictive data mining technique on the same dataset and the outcome reveals that Decision Tree out performs and sometime Bayesian classification is having similar accuracy as of decision tree but other predictive methods like KNN, Neural Networks, Classification based on clustering are not performing well. The second conclusion is that the accuracy of the Decision Tree and Bayesian Classification further improves after applying genetic algorithm to reduce the actual data size to get the optimal subset of attribute sufficient for heart disease prediction.

Himigiri this research intends to provide a survey of current techniques of knowledge discovery in databases using data mining techniques that are in use in today's medical research particularly in heart disease prediction. Number of experiment has been conducted to compare the performance of predictive data mining technique on the same dataset and the outcome reveals that Decision Tree outperforms and sometime Bayesian classification is having similar accuracy as of decision tree.

Artificially Intelligent (AI) powered tools are able to deal with uncertain and incomplete datasets (Shadabi and Sharma, 2008). Neural network classifiers have been successfully used for prediction purposes in many complex situations. Research demonstrates that AI-based data mining tools have been also successfully used inmany medical environments. This research advances the understanding of the application of artificial intelligence and data mining tools to clinical data by demonstrating the potential of these techniques incomplex clinical situations.

Joshi and Jain (2010) recently, different works proposed a new way to mine patterns in transposed databases where a database with thousands of attributes but only tens of objects. In this case, mining the transposed database runs through a smaller search space. In this research, they systematically explore the search space off requent patterns mining and represent database in transposed form. They develop an algorithm (termed DFPMT-a dynamic approach for frequent patterns mining using transposition of database) for mining frequent patterns which are based on Apriori algorithm and used dynamic approach like longest common subsequence. The main distinguishing factors among the proposed schemes is the database stores in transposed form and in each iteration database is filter/reduce by generating LCS of transaction ID for each pattern. Their solutions provide faster result. A quantitative exploration of these tradeoffs is conducted through an extensive experimental study on synthetic and real-life datasets.

Peter and Somasundaram (2012) in this research, the use of pattern recognition and data mining techniques into risk prediction models in the clinical domain of cardiovascular medicine is proposed. The data is to be modeled and classified by using classification data mining technique. Some of the limitations of the conventional medical scoring systems are that there is a presence of intrinsic linear combinations of variables in the input set and hence they are not adept at modeling non-linear complex interactions in medical domains. This limitation is handled in this research by use of classification models which can implicitly detect complex non-linear relationships between dependent and independent variables as well as the ability to detect all possible interactions between predictor variables (Wang and Miao, 2012).

The system proposed in this research uses this vast storage of information so that diagnosis based on this historical data can be made (Isola *et al.*, 2012). It focuses on computing the probability of occurrence of a particular ailment from the medical data by mining it using a unique algorithm which increases accuracy of such diagnosis by combining the key points of Neural Networks, Large Memory Storage and Retrieval (LAMSTAR), k-NN and Differential Diagnosis all integrated into one single algorithm.

Data cleaning: Real world data tends to be incomplete, noisy and inconsistent. Data cleaning routine attempt to fill in the missing values, smooth out noise while identifying outliers and correct inconsistencies in the data.

Missing values: Various strategies were applied to generate the missing values depending on the importance of the missing value and its relation to the search domain. Either fill in the missing value manually or use aglobal constant to fill in the missing value.

GENERAL DESCRIPTION OF OUR MEDICAL DATA

The medical dataset we are mining describes the proles of patients being treated for coronary Heart disease. All medical information is put in one_le having several records. Each record corresponds to the most relevant information of one.

RESULTS

Simulation tool:

- MATLAB editor is used for writing the code to implement our algorithm
- The result will be shown in the command window of MATLAB

Objective of present work:

- The main objective of the work is to define new two layered coronary heart disease prediction system. The work will depend on the multi parameters
- A neuro-fuzzy approach will be implemented on two layers
- Each layer consists of different parameters with first layer having critical ones and second layer having the rest
- The work is about to identify the disease chances accurately (Fig. 1)

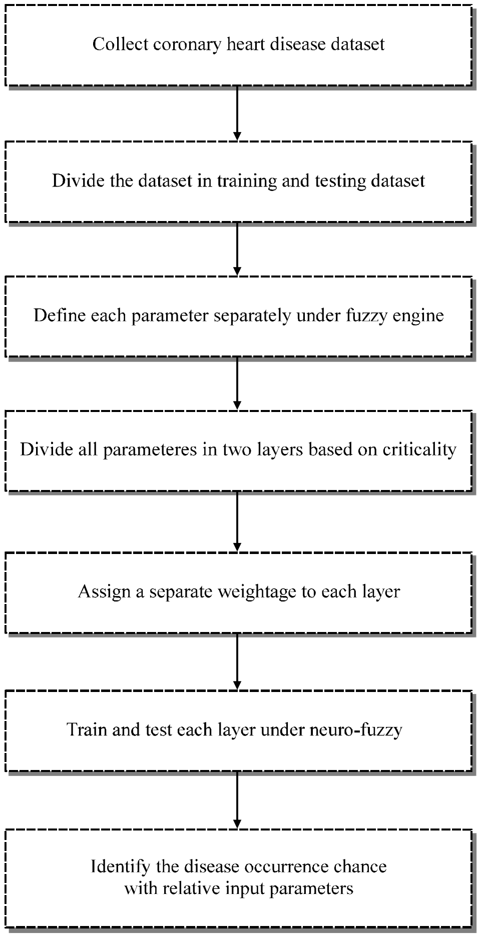
Collect coronary heart disease dataset
Divide the dataset in training and testing dataset
Define each parameter separately under fuzzy engine
Divide all parameteres in two layers based on criticality
Assign a separate weightage to each layer
Train and test each layer under neuro-fuzzy
Identify the disease occurrence chance
with relative input parameters

Fig. 1: Flow chart of present work

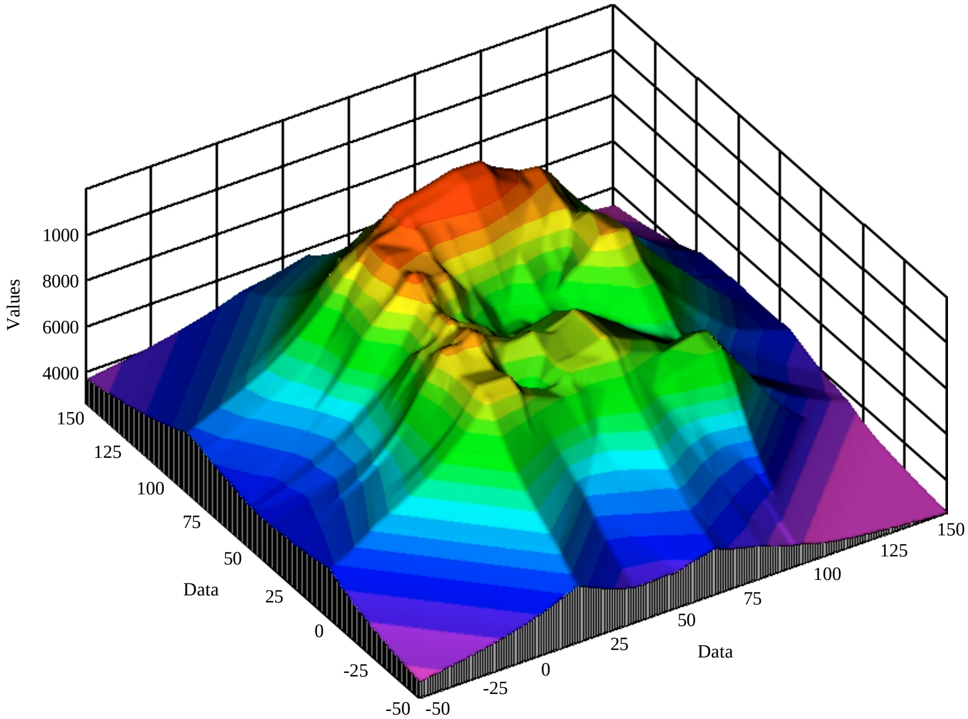

Fig. 2: Training input data

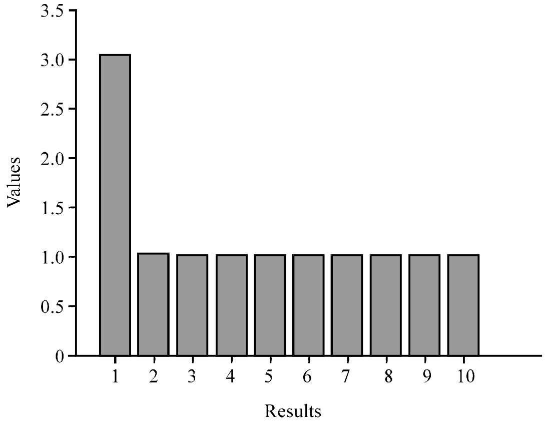

Fig. 3: Predictive result driven

TRAINING INPUT DATA

In this Fig. 2, the input to the current system is perform is performed called the training input. The training data is taken under different medical parameters such as age, cholesterol, blood pressure, etc. The available UCI dataset is divided in two parts called training dataset and the testing dataset. We have selected 80% data values as the training dataset and 20% dataset as the testing dataset. Back Propagation algorithm is used to train artificial neural network. In the Fig. 2, the patient analysis based on training dataset is defined. The first image in Fig. 1 is showing the increment respective to the training dataset in on different medical parameters and the second image is showing the error effectiveness.

PREDICTIVE RESULT DRIVEN

Figure 3 is showing the predictive results driven from the model. As we can see, the model has shown the incremental change in the patient information under different parameters.

CONCLUSION

Clinical medicine is one of the most interesting areas in which data mining may have an important practical impact. The widespread availability of large clinical data collections enables thorough retrospective analysis which may give health care institutions an unprecedented opportunity to better understand the nature and peculiarity of the under going clinical processes. In present work, we have designed a system to identify the chances of a coronary heart disease. We have divided all parameters in two levels according to criticality and each level is assigned separate weightage. Finally, both levels are considered to derive a final decision. We have implemented neuro-fuzzy integrated approach at two levels. So, error rate is very low and work efficiency is high. In this research, we have performed the analysis for coronary heart disease. In future, we can use the same neuro-fuzzy integrated approach to perform the analysis on some other disease.

REFERENCES

Isola, R., R. Carvalho and A.K. Tripathy, 2012. Knowledge Discovery in Medical Systems Using Differential Diagnosis, LAMSTAR and NN. Inf. Technol. Biomed., 16: 1287-1295.

Joshi, S. and R.C. Jain, 2010. A dynamic approach for frequent pattern mining using transposition of database. Proceedings of the Second International Conference on Communication Software and Networks, February 26-28, 2010, Singapore, pp: 498-501.

Koutsojannis, C. and I. Hatzilygeroudis, 2007. Using a Neurofuzzy Approach in a Medical Application. In: Knowledge-Based Intelligent Information and Engineering Systems. Apolloni, B., J. Robert and J. Lakhmi (Eds.). Springer, Berlin Heidelberg, Germany, ISBN: 978-3-540-74826-7, pp: 477-484.

Lin, Z.K., W.G. Yi, M.Y. Lu, Z. Liu and H. Xu, 2009. Correlation research of association rules and application in the data about coronary heart disease. Proceedings of the International Conference of Soft Computing and Pattern Recognition, December 4-7, 2009, Malacca, pp: 143-148.

Patel, S.B., P.K. Yadav and D.P. Shukla, 2013. Predict the diagnosis of heart disease patients using classification mining techniques. IOSR J. Agric. Vet. Sci., 4: 61-64.

Peter, T.J. and K. Somasundaram, 2012. An empirical study on prediction of heart disease using classification data mining techniques. Proceedings of the International Conference on Advances in Engineering, Science and Management, March 30-31, 2012, Nagapattinam, Tamil Nadu, pp: 514-518.

Shadabi, F. and D. Sharma, 2008. Artificial intelligence and data mining techniques in medicine-success stories. Proceedings of the International Conference on IEEE BioMedical Engineering and Informatics, May 27-30, 2008, Sanya, pp: 235-239.

Soni, J., U. Ansari, D. Sharma and S. Soni, 2011. Predictive data mining for medical diagnosis: An overview of heart disease prediction. Int. J. Comput. Appl., 17: 43-48.

Wang, Z. and M. Miao, 2012. Discovery the relationship in properties of traditional Chinese medicine based on data mining. Proceedings of the International Symposium on Information Technology in Medicine and Education, August, 3-5, 2012, Hokodate, Hokkaido, pp: 782-785.