Kauser Ahmed P /J. Pharm. Sci. & Res. Vol. 9(10), 2017, 1886-1888

ISSN:0975-1459
Journal of Pharmaceutical Sciences and Research
www.jpsr.pharmainfo.in

Analysis of Data Mining Tools for Disease Prediction

Kauser Ahmed P

School of Computer Science and Engineering,
VIT University, Vellore-632014, Tamil Nadu, India

Abstract

Adaptation of information technology has leads to creation of several applications in health care informatics. Health care informatics is generating large amount of data. These data can be processed using data mining techniques to predict the diseases. Data mining is the process of analyzing, extracting data and furnishes the data as knowledge which forms the relationship within the available data. Some of the data mining techniques include association, clustering, classification and prediction. Various data mining tools are compared to analyze the performance of health care data and disease prediction.

Keywords: Clustering, Classification, Data mining tools, Disease prediction, Health care.

INTRODUCTION

Significant advances in information technology results in excessive growth of data in health care informatics [1]. Health care informatics data includes hospital details, patient's details, disease details and treatment cost. These huge data are generated from different sources and format. It can have irrelevant attributes and missing data. Applying data mining techniques is a key approach to extract knowledge from large disease data. Data mining has various methods to extract knowledge from huge disease data set. Data mining techniques like classification, clustering and rule mining can be used to analyze data and extract meaningful information. Some of the important current applications of data mining in health care includes predicting the future outcomes of diseases based on previous data collected from similar diseases, diagnosis of disease based on patient data, analyzing treatment costs and demand of resources, preprocessing of noisy, missing data and minimizing the time to wait for the disease diagnosis. Data mining tools like Weka, Rapid miner and Orange [2, 3, 4] are used to analyze and predict better result for health care data. New and current data mining tools and technologies are used in disease diagnosis and health care informatics to improve the health care services in cost effective manner and minimizing the time for disease diagnosis.

The organization of this paper is as follows. Section 1 describes the fundamentals of data mining. Section 2 list out various data mining techniques used in health care. Data mining tools are discussed in section 3. Results and discussion are highlighted in section 4. Concluding remarks are given in section 5.

1. FUNDAMENTALS OF DATA MINING

Data mining is concerned with the process of computationally extracting unknown knowledge from huge sets of data. Extraction of useful knowledge from the enormous data sets and providing decision-making results for the diagnosis and treatment of diseases is very important. Data mining can be used to extract knowledge by analyzing and predicting various diseases. Health care data mining has great potential to discover the hidden patterns in the data sets of the medical domain. Various data mining techniques are available with their suitability dependent on the health care data. Data mining applications in health care can have a wonderful potential and effectiveness. It automates the process of finding predictive information in huge databases.

Disease prediction plays an important role in data mining. Finding of a disease requires the performance of a number of tests on the patient. However, use of data mining techniques, can reduce the number of tests. This reduced test set plays significant role in performance and time. Health care data mining is an important task because it allows doctors to see which attributes are more important for diagnosis such as age, weight, symptoms etc. This will help the doctors diagnose the disease more efficiently.

Knowledge discovery in databases is the process of finding useful information and patterns in data. Knowledge discovery in databases can be done using data mining. It uses algorithms to extract the information and patterns derived by the knowledge discovery in databases process. Various stages of knowledge discovery in databases process is highlighted in Fig.1.

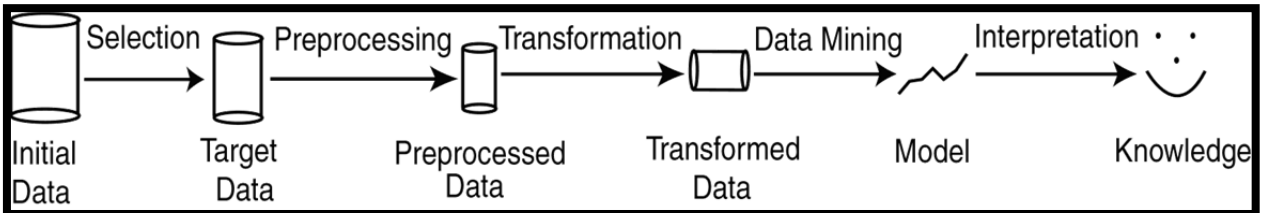

Fig.1: Healthcare Knowledge Discovery process

Various stages of knowledge discovery in databases process is describe as follows. In Selection stage, it obtains the data different resources. In preprocessing stage, it removes the unwanted missing and noisy data and furnished the clean data which can format to a common format in transformation stage. Then data mining techniques is applied to get desired output. Finally in the in the interpretation stage, it will present the result to end user in a meaningful manner.

2. DATA MINING TECHNIQUES

Data mining techniques like classification, clustering and association rules are widely used in disease data analysis.

Classification

Classification is a machine learning based data mining technique. Classification is used to classify each information in a set of data into one of predefined set of groups or classes. It makes use mathematical techniques such as decision trees, linear programming, neural network and statistics to classify the data into different groups. Modern classification techniques provide more intelligent methods for effective prediction of diseases [5]. Different types of classification techniques includes Support vector machine, discriminant analysis, naive based, decision trees, linear and non linear regression.

Clustering

Clustering is a data mining technique that makes cluster of objects that have similar characteristic using automatic technique. Clustering technique defines the classes and put objects in them where the class is not predefined. Different types of cluster techniques includes K-means, Fuzzy C-means (FCM), Rough C-means (RCM), Rough-Fuzzy C-means (RFCM) , Robust RFCM (rRFCM), hierarchical and Gaussian mixture.

Association rule mining

Association rule learning is a popular and well researched method for finding interesting relations between different data in large databases. It is intended to identify well built rules discovered in databases using different procedures of importance based on input data set. Association rule mining is the data mining process of finding the rules, finding frequent patterns, associations, correlations, or causal structures among sets of items that may govern associations and causal objects between sets of items. Understand customer buying habits by finding associations and correlations between the different items that customers place in their “shopping basket”. The main applications of association rule mining includes basket data analysis, cross-marketing and catalog design. The above data mining techniques can be used in the diagnosis of diseases [6].

3. DATA MINING TOOLS

Data mining tools like Weka, Rapidminer, Orange and Knime are used to perform various data mining techniques.

WEKA

The Waikato Environment for Knowledge Analysis (WEKA) [7] is an open source software and machine learning toolkit introduced by Waikato University, New Zealand. WEKA supports several standard data mining tasks like data preprocessing, clustering, classification, regression, visualization and feature selection New algorithms can also be implemented using WEKA with existing data mining and machine learning techniques. WEKA provides various sources for loading data, including files, URLs and databases. It supports file formats include WEKA"s own ARFF format, CSV, Lib SVMs format, and C4.5’s format. Many evaluation criteria are also provided in WEKA such as confusion matrix, precision, recall, true positive and false negative, etc. Some of the advantages of WEKA tool includes Open source, platform independent and portable, graphical user interface and contains very large collection of different data mining algorithms.

RAPIDMINER

RAPIDMINER (RM) [8] is open source software which provides a good environment for data mining processes. It has the facility of drag-and-drop which is used to construct the dataflow. It support different file formats. Regression, classification and clustering tasks can be performed easily with different learning algorithms. Rapid Miner supports a large number of the classification and regression algorithms, decision trees, association rules, clustering algorithms, and many features are available for data pre-processing, normalization, filtering and data analysis. It can import data from different traditional and standard databases.

ORANGE

ORANGE [9] is an open source data mining tool developed at the Bioinformatics Laboratory at the University of Ljubljana. Applications can be implemented using scripting and visual programming. Python library is available for data manipulation and widget alteration. Programming is performed by placing widgets on the canvas and connecting their inputs and outputs. This tool is suitable for machine learning and data mining algorithms. It can be easily used by both researchers of data mining and inexperienced users who want to develop and test their own algorithms. It gives advantage of reusing as much of the code as possible.

KNIME

KNIME (Konstanz Information Miner) [10] is a general purpose open source data mining tool developed and maintained by the Swiss company. It is implemented on the Eclipse platform and has facility of data integration, processing, exploration, and analysis platform. KNIME can be integrated with other data mining tools such as R and WEKA.

4. RESULTS AND DISCUSSION

Three data mining tools such as Weka, Orange and Rapidminer are analyzed and their performances are compared on Iris data set [11] using Naive bayes classification algorithm. The measure taken for comparing these data mining tools is classification accuracy based on precision and recall. The classification accuracy of data mining tools on Iris data set is highlighted in Table 1 and Fig.2.

Table 1: Classification Accuracy of data mining tools on Iris data set

Classification Technique	Tools	Precision	Recall
Naïve Bayes	Orange	0.857	0.857
	Weka	0.96	0.96
	Rapidminer	0.95	0.95

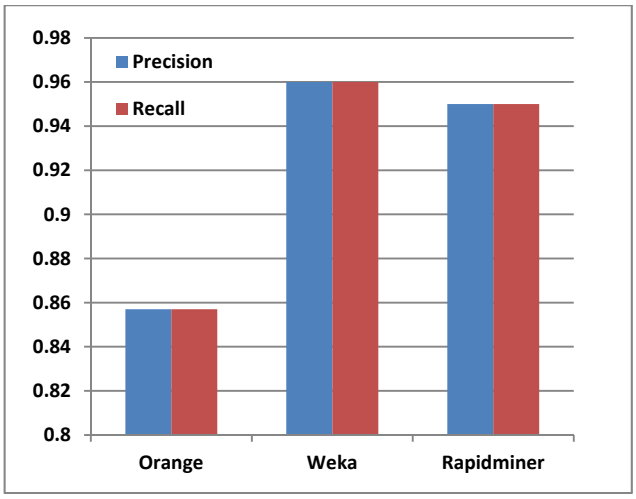

Fig.2: Classification Accuracy of data mining tools on Iris data set.

Based on analysis the following result has been derived from different data mining tools.

WEKA is the best tool for a beginner since it contains many in-built and experimental features and no prior knowledge of coding is required.

RapidMiner is the only tool which is independent of language limitations and has statistical and predictive analytical capabilities.

ORANGE and RapidMiner in comparison are the tools that are for advanced users since it requires advanced knowledge in coding.

5. CONCLUSION

Data mining techniques helps in finding the hidden knowledge in a group of disease data that can be used to analyze and predict the future behavior of diseases. Classification is one the data mining techniques which assigned a class label to a set of unclassified cases. The main objective of this paper is to compare the data mining tools on the basis of their classification accuracy. According to the result of three data mining tools used in this paper, it has been observed that different data mining tools are furnishing different results on same data set with different classification algorithm. WEKA is showing best classification accuracy when compared to rapidminer and orange. In future, more disease dataset can be used for classification techniques and other data mining techniques such as clustering can be used to compare the performance of various data mining tools.

REFERENCES

1. Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine learning and data mining methods in diabetes research. Computational and structural biotechnology journal. 2017; 104-116.
2. Patil PH, Thube S, Ratnaparkhi B, Rajeswari K. Analysis of Different Data Mining Tools using Classification, Clustering and Association Rule Mining. International Journal of Computer Applications. 2014; 93(8):35-39.
3. Usha Rani D. Survey on Data Mining Tools and Techniques in Medical Field. International Journal of Advanced Networking & Applications. 2017; 8(5):51-54.
4. Devi SK, Krishnapriya S, Kalita D. Prediction of Heart Disease using Data Mining Techniques. Indian Journal of Science and Technology. 2016;9(39):1-5.
5. Bhatla N, Jyoti K. An analysis of heart disease prediction using different data mining techniques. International Journal of Engineering. 2012;1(8):1-4.
6. Fatima M, Pasha M. Survey of Machine Learning Algorithms for Disease Diagnostic. Journal of Intelligent Learning Systems and Applications. 2017;9(1):1-16.
7. http://www.cs.waikato.ac.nz/ml/weka/ [Last accessed: 22/08/2017].
8. https://rapidminer.com/ [Last accessed: 22/08/2017].
9. https://orange.biolab.si/ [Last accessed: 22/08/2017].
10. https://www.knime.com/ [Last accessed: 22/08/2017].
11. https://archive.ics.uci.edu/ml/datasets/iris [Last accessed: 22/08/2017].