IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727
PP 04-07
www.iosrjournals.org

Data Mining Applications in Healthcare

Arun Pushpan[1], Ali Akbar N[2]

[1]Mtech Scholar Department of Computer Science Engineering Govt Engineering College Mananthavady, Wayanad Kerala,

[2]Associate Professor Department of Computer Science Engineering Govt Engineering College Mananthavady, Wayanad Kerala, India

Abstract: *Data mining provides a variety of latest method for data analysis and discover new useful knowledge.In different research areas,data mining is gaining popularity due to its infinite methodologies and applications to mine the information. Data mining techniques plays a vital role for uncovering new trends in healthcare organization which is also for all the parties associated with this field.Usage of such data mining techniques on medical data determine useful trends and patterns that are used in analysis and decision making. This survey features various data mining techniques such as classification, clustering, association, regression in health domain. It can find out some useful knowledge from large databases. Now-a-days large volume of data is being collected and stored at high speed. Traditional data analysis techniques have limitations. Human analyst may take time to discover useful information and much of the data is never analyzed at all. Automated analysis of massive data sets is considered as data mining. Compared with other data mining areas, medical data mining has some unique characteristics [1]. Since medical files are related to human subjects, privacy concern is taken more seriously than other data mining tasks.*

I. Introduction

Data mining is an algorithmic techniques for extracting new useful patterns from collected raw data .Today, huge amount of data is produced by healthcare industries. The data produced includes data about hospitals, resources,electronic patient records, disease diagnosis , etc. Data mining pro-cesses includes the hypothesis framing ,data gathering , pre-processing,model estimation,model understanding and then finding the conclusions. Healthcare industries have data that clutches complex information about the patients and their medical conditions. Data mining is becoming popular in various research arenas due to its wide variety of applications and system of methods used to mine informations that are useful in correct manner. Data mining techniques have the ability to detect hidden patterns or relationships among the objects in medical records. Data mining techniques have been applied on medical data during the last few decades. This is for determining useful trends or patterns that are used in analysis and making proper decisions. The infinite potential of data mining is utilized for predicting the different kind of diseases more efficiently and effectually in health care data . Some data mining methods used in medical field includes Association, Clustering, Classification, etc.

Knowledge Discovery (KDD) and Data Mining are two related terms and they are used interchangeably. Knowledge Discovery in database is systematized in different stages.The first stage is selection of data in which data is collected from different sources, the second stage is preprocessing the data that is collected, the third stage is transformation of data into suitable format for their further processing, the fourth stage includes data mining where suitable Data Mining technique is applied on the transformed data in order to extract valuable information and Interpretation is the final stage. The process of retrieving high-level knowledge from base level data can be considered as knowledge discovery in databases. It is an iterative process that consist of steps like Selecting valid data, Preprocessing the selected data, Conversion of data into an understandable format, Data mining to extract useful informa-tion and Interpretation/Evaluation of data. In order to perform processing the selection step collects the heterogeneous data from different sources. The medical records available in real life may be incomplete, complex, noisy, inconsistent, and irrelevant which requires a selection process. Selection process gathers the most important data from which knowledge is to be derived. Basic operations are performed by preprocessing step. It includes the elimination of noisy data, try to finding out the unavailable data or to develop a policy for handling those data, detect or remove outliers and resolution of inconsistencies among the data is performed. Transformation step transforms the data. It convert the data into a format which is apt for extraction by performing task like smoothing, generalization, normalization, discretization and aggregation. Data reduction task performs data shrinking and represents the same data in to less volume, but produces the similar analytical outcomes [2]. Data mining is an important component in KDD process. It involves selecting the appropriate data mining algorithms and using those algorithms for generating previously unknown and hypothetically useful information from the data that are stored in the database. This comprises deciding which models or algorithms and which parameters may be suitable and comparing a specific data mining method with the general standards of the KDD process. Regression, Summarization, Classification, clustering etc. are the general data mining techniques. Interpretation or Evaluation step present the mined

patterns in understandable format. The type of data representa-tion changes with different types of information needs, in this step the interpretation of mined patterns occur. Evaluation of the results is prepared with significance testing and statistical justification [1].

A. Data Mining Techniques

In data mining there are mainly two types of learning techniques. The two approaches are supervised learning and unsupervised learning.

1) Supervised Learning Techniques: Supervised Learning is the most common method used for learning purpose. Some predefined models are used for training purpose. Using the training data set the model is build. A new incoming data will be checked against the trained model and determine the class label of the new data. The disadvantages of supervised learning techniques include the difficulty in gathering class labels. If there is bulk input then it becomes expensive to label . Classification and Regression methods come under this category.
2) Unsupervised Learning Techniques: Unlike supervised data-mining methods, in unsupervised methods, no result will get from its surroundings. Although the visualization of how a machine can be trained without any response from its sur-roundings is difficult, these methods work well. It is very likely to build a proper model for unsupervised learning methods that support on the idea that the mechanisms aim is to use input characterization to predict prospective input, decision making, effectively communicating the input to another mechanism, and so on. Unsupervised learning methods can find patterns from a collection of data which can also be unstructured noise. The general unsupervised learning methods are dimensionality reduction and clustering. The main benefit of using supervised techniques over unsupervised is that once the classifier has

been trained, it can be easily utilized on any same kind of datasets. Association rule mining and Clustering are the examples of unsupervised techniques.

Three of the most widely used data mining algorithms in health care are classification method, clustering and associa-tion rule mining.

B. Classification

Class label of a categorical attribute can be predicted by using classification methods. Class is the dependent attribute in which users are most interested. Using the set of independent attributes the value of this dependent attribute is predicted. In the medical field, based on the symptoms and health conditions of the patient, classification method can be used to define the diagnosis procedure and prognosis prediction [1][3]. Classification method consists of mainly two steps. They are labeled as classification step and learning step. Learning step takes the input as training data and it builds a classifier that generates the classification rules. In classification step, the accuracy of the classifier is tested using test data . Predictive accuracy of the classifier is estimated. The percentage of test data whose class labels are correctly classified by the classifier is the accuracy of the classifier. If the accuracy of the classifier is considered acceptable then it is used to classify the class labels of future data tuples.

The relevant and irrelevant attributes is identified before applying classification algorithm to a data set. For example age and date of birth are relevant attributes. It is important to remove attributes from datasets that can be derived from one another. Only one attribute among them is get eliminated. The totally irrelevant attributes should be identified. These attributes may act as noise and slow down the efficiency and accuracy of classification method. For example gender attribute on prostate cancer data mining is irrelevant. High prediction power is the advantage of classification methods. That is why it is preferred in medical data mining. following are the classification models of data mining technique.

Decision Trees Neural Networks Naive Bayes Classifier

1) Decision Trees: Decision tree approach is considered as an important classification technique in data mining. Tree structured models are build by using this method. The tree like structure is build by dividing the large dataset into number of small sets. Finally an associated tree is constructed. One of the advantage with decision tree is that it is able to handle both categorical information and numerical data. This method takes a decision based on already determined order in different attributes for medical purpose. Iterative Dichotomized (ID3) is the most commonly used decision tree algorithm. ID3 classifies data in tree structure by using information entropy and information gain [3][6][7].

steps involved in ID3 algorithm

Build classification attributes from the dataset Now compute classification entropy

If entropy value is higher, then the possibility to enhance classification is also higher.

For each attribute, compute information gain utilizing classification attribute.

The attribute with high information gain value is used to divide the set on particular iteration.

Remove node attribute

repeat above steps until all attributes have been utilized, or the same classification values stays in rows of minimized set.

The advantages of ID3 includes:

It is easy to interpret and easy to understand Rules can be easily generated

Allows the addition of new data

The disadvantages of ID3 includes:

Difficult to handle non numeric data It is very time consuming

Sometimes it is difficult to understand trees with many branches

2) Neural Networks: It is an information processing method. It consist of huge number of highly interconnected elements on the layers. It consist of three layers, namely, input layer, output layer and hidden layer. Input layer is the principal layer and final layer is the output layer. The layer in between input and output layer is the hidden layer. An output layer is formed when an input layer is communicating with one or more hidden layers. Commonly it consist of some learning rules that can modify the connection weight. This learning can be supervised learning or unsupervised learning [4][7].

The advantages of neural networks include:

it can easily handle error data or missing data No reprogram is needed if it is trained once

It is able to work with large number of datasets.

The disadvantages include:

It needs training to work well

Sometimes it needs high processing time for large net-works

It is not possible to retrain the neural networks, that is no modified data can be added into an existing network.

3) Naive Bayes Classifier: Naive Bayes Classifier is simple classifier with independence assumptions based on Bayes theorem. It also known as independent feature model. It can build models having the capability to predict something. It is a supervised learning method. Naive Bayes Classifier assumes that the presence or absence of a particular feature of a class is unrelated to the presence or absence of other feature. This property of Naive Bayes Classifier is known as class conditional independence. It can be used for medical diagnosis [5][6].

Advantages of neural networks include:

The parameters can easily found with a small volume of training set

Can easily handle huge volume of data Will avoid irrelevant features

Training and classification will be faster It will work with discrete and real data

Disadvantages of neural networks include:

Less accurate

May assume features that are independent

The dependency among variables may not be handled by the classifier

C. Weighted association classifier

Here classification is based on Weighted Association Rule. Weighted Association Rule (WARM) mining uses confidence framework and weighted support to derive association rule from dataset. Major steps in WAC are as follows.Initial preprocessing step to convert the data in format apt for mining Each attribute is provided with a weight of value 0 to1. This value indicates its importance in prediction. If an attribute have a value equal t0 0.9, then its impact will be high. Similarly if the value is 0.1, then its impact is low.

WARM is applied to generate patterns. Rules generated here are known as Classification Association Rule (CAR).

Store these rules in Rule Base.

CAR rule from the Rule Base is applied to new data whenever it is provided

D. Clustering

A cluster can be considered as a group consisting of similar elements. All the elements in the same cluster shows similarity with each other and dissimilarity with elements with other cluster elements. Cluster analysis aims to group data elements having similar properties and characteristics together and forms clusters. Clustering is an unsupervised learning process because it occurs by observing only independent attributes in the dataset. Clustering doesn’t consider the concept of class. Therefore it can be applied for the studies that contains huge amount of data. In this case only less information will be known about the data. High intra-class similarity and low inter-class similarity are the main properties of a good cluster. Quality of the cluster is determined by the method of the similarity measure used. The ability to find out hidden patterns and its measure determines the quality of clustering technique.

Clustering can be used when no information about data is known or very little information is known about the data. Thus, to study about micro array data sequences, clustering can be used. Only numerical attributes can be handled by almost every clustering algorithm. However, most health care data bases have number of attributes of categorical type. Numeric conversion of such attributes is also possible, such conversion may distorts the distance between categories. There are some algorithms that handles categorical attributes without converting them into numerical form. For example Farthest First and Two-step Cluster analysis methods can handle the categorical data attributes [2].

E. Association Rule Mining

Association rule mining also known as frequent item set or pattern mining. It is often employed to find frequent patterns, associations, relationships among set of objects in the dataset. In the case of health care industry, one can use this association method to discover relationship among disease symptoms, relationship among health conditions of various patients suffering from same disease and relationship among various diseases. Researchers can derive evidence-based hy-potheses about health conditions and symptoms contributing to a disease or complications. Association rule mining algorithms are not evaluated based on its accuracy.It is because every association algorithm mines all association rules. Evaluation of association mining algorithms are based on its efficiency to mine hidden rules from large dataset.

Classification algorithm is mainly focused towards class label whereas association mining algorithms are mainly used to discover relationship among all attributes. In the case of good association mining method,it should ignore association rules which are of meaningless [3].

II. Conclusion

Data mining support the medical experts to discover new life saving information. The medical institutions apply this on their existing data base and can uncover latest, valid and useful knowledges . More over this data mining helps in better med-ical policies like strict sterilization and sanitation, vaccination planning etc. As there is voluminous records in this industry and because of this, it has become requisite to use data mining techniques to help in decision support and prediction in the field of healthcare to identify the kind of disease. The medical data mining produces business intelligence which is useful for diagnosing of the disease. This paper observe some data mining techniques that has been employed for medical data. This paper throws light into data mining techniques that is used for medical data for various diseases which are identified and diagnosed for human health.

References

[1]. Sarvnaz Karimi, Chen Wang, Alejandro Metke jimenez, Raj Gaire, Text and Data Mining Techniques in Adverse Drug Reaction Detection, ACM, 2015.

[2]. Hian Chye Koh, Gerald Tan, Alejandro, Data Mining Applications in Healthcare, Journal of Healthcare Information Management Vol. 19, No. 2.

[3]. Marina Evrim Johnson, Nagen Nagarur, Multi-stage methodology to detect health insurance claim fraud , Springer, 2015.

[4]. Syed Umar Amin, Kavita Agarwal, Dr. Rizwan Beg, Genetic Neural Network Based Data Mining in Prediction of Heart Disease Using Risk Factors, Proceedings of 2013 IEEE Conference on Information and Communication Technologies, 2013.

[5]. Dr. Shailendra Narayan Singh, Monika Gandhi, Predictions in Heart Disease Using Techniques of Data Mining, 2015 1st International Conference on Futuristic trend in Computational Analysis and Knowledge Management,IEEE, 2015.

[6]. Sujata Joshi and Mydhili K. Nair, Prediction of Heart Disease Using Classification Based Data Mining Techniques, Computational Intelli-gence in Data Mining - Volume 2, Springer, 2015.

[7]. Ritika Chadha, Shubhankar Mayank, Anurag Vardhan and Tribikram Pradhan, Application of Data Mining Techniques on Heart Disease Prediction: A Survey, Emerging Research in Computing, Information, Communication and Applications, Springer, 2016.