International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 03 | Mar-2016 www.irjet.net p-ISSN: 2395-0072

A STUDY ON DATA MINING PREDICTION TECHNIQUES IN HEALTHCARE SECTOR

Dr.B.Srinivasan [1], K.Pavya [2]

[1]Department of Computer Science, Gobi Arts & Science College, Bharathiar University, Tamil Nadu, India [2]Department of Computer Science, Vellalar College for Women, Bharathiar University, Tamil Nadu, India

---

Abstract: *One of the fastest growing fields is health care industries. The medical industries have huge amount of data set collections about patient details, diagnosis and medications. To turns these data is into useful pattern and to predicting comingup trends data mining approaches are used in health care industries. The healthcare industry collects huge amount of healthcare data which are not "mined" to find out hidden information. The medical industries come crossways with new treatments and medicine every day. The healthcare industries should provide better diagnosis and therapy to the patients to attain good quality of service. This paper explores different data mining techniques which are used in medicine field for good decision making.*

Key Words: Data mining, KDD, Prediction techniques, Decision making.

I. INTRODUCTION

Data mining is the method for finding unknown values from enormous amount of data. As the patients population increases the medical databases also increasing every day. The transactions and investigation of these medical data is difficult without the computer based analysis system. The computer based analysis system indicates the mechanized medical diagnosis system. This mechanized diagnosis system support the medical practitioner to make good decision in treatment and disease. Data mining is the huge areas for the doctors to handling the huge amount of patient's data sets in many ways such as make sense of complex diagnostic tests, interpreting previous results, and combining the dissimilar data together. Traditionally infirmary decision is shaped by the medical practitioner's observations and fore knowledge rather than the knowledge which obtain from the large amount of data. This mechanized diagnosis system leads to increases the quality of service provided to the patients and decreases the medical costs.

II. KNOWLEDGE DISCOVERY AND DATA MINING

This section provides an introduction to knowledge discovery and data mining.

A. Knowledge Discovery Process

The terms Knowledge Discovery in Databases (KDD) and Data Mining are frequently used interchangeably. KDD is the process of changing the low-level data into high-level knowledge. Hence, KDD refers to the nontrivial removal of implicit, previously unknown and potentially useful information from data in databases. While data mining and KDD are often treated as comparable words but in real data mining is an essential step in the KDD process.

The Knowledge Discovery in Databases process comprise of a few steps leading from raw data collections to some form of new information. The iterative process consists of the following steps:

(1) *Data cleaning:* also known as data cleansing it is a phase in which noise data and unrelated data are removed from the collection.

(2) *Data integration*: at this stage, several data sources, often heterogeneous, may be shared in a common source.

(3) *Data selection:* at this step, the data related to the analysis is decided on and retrieve from the data collection.

(4) *Data transformation:* also known as data consolidation, it is a phase in which the chosen data is transformed into forms appropriate for the mining procedure.

(5) *Data mining:* it is the essential step in which clever techniques are applied to extract patterns potentially useful.

(6) *Pattern evaluation:* this step, firmly interesting patterns representing knowledge are known based on given measures.

(7) *Knowledge representation:* is the last phase in which the discovered knowledge is visually represented to the

user. In this pace visualization techniques are used to help users understand and interpret the data mining results.

The following figure.1 shows data mining as a step in an iterative knowledge discovery process.

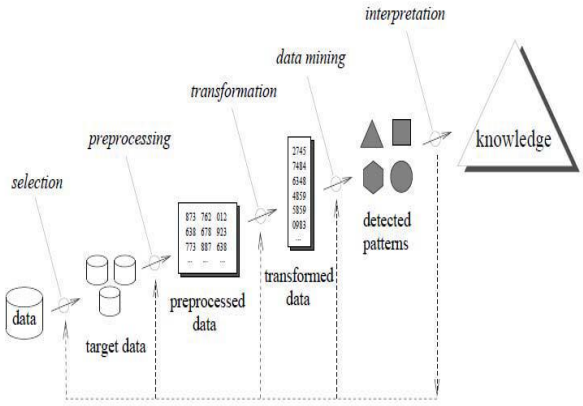

Figure1. Steps in KDD

B. Data Mining Process

In the KDD process, the data mining methods are for extracting patterns from data. The patterns that can be exposed depend upon the data mining tasks applied. Generally, there are two types of data mining tasks:
descriptive data mining tasks that explain the general properties of the existing data, and *predictive data mining tasks* that attempt to do predictions based on available data. Data mining can be done on data which are in textual, quantitative or multimedia forms. Data mining applications can use dissimilar kind of parameters to observe the data. They include association (patterns where one event is related to another event), sequence or path analysis (patterns where one event leads to another event), classification (identification of new patterns with predefined targets) and clustering (grouping of identical or alike objects).Data mining involves some of the following key steps:
(1) *Problem definition:* The first step is to discover goals. Based on the defined goal, the correct series of tools can be applied to the data to build the corresponding behavioral model.
(2) *Data exploration:* If the value of data is not suitable for an perfect model then recommendations on future data collection and storage strategies can be made at this. For analysis, all data needs to be consolidated so that it can be treated consistently.
(3) *Data preparation:* The purpose of this step is to clean and convert the data so that missing and invalid values are treated and all known valid values are made reliable for more robust analysis.
(4) *Modeling:* Based on the data and the desired outcomes, a data mining algorithm or group of algorithms is selected for analysis. These algorithms include classical techniques such as statistics, neighborhoods and clustering but also next invention techniques such as decision trees, networks and rule based algorithms. The specific algorithm is selected based on the particular objective to be achieved and the quality of the data to be analyzed.
(5) *Evaluation and Deployment:* Based on the outcome of the data mining algorithms, an analysis is conducted to find out key conclusions from the analysis and create a sequence of recommendations for consideration.

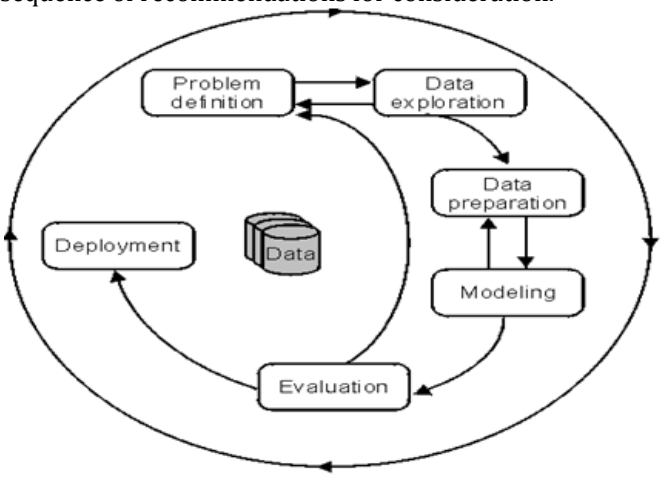

Figure 2. Data Mining Process Representation

III. HEALTHCARE DATA MINING

The increasing research area in data mining technology is Healthcare data mining. Data mining holds immense promising for healthcare management to allow health system to systematically use data and analysis to progress the care and decrease the cost concurrently could apply to as much as 30% of overall healthcare spending. In the healthcare managing data mining prediction are playing vigorous role. Some of the prediction based data mining techniques are as follows:

1. Neural network
2. Bayesian Classifiers
3. Decision tree
4. Support Vector Machine

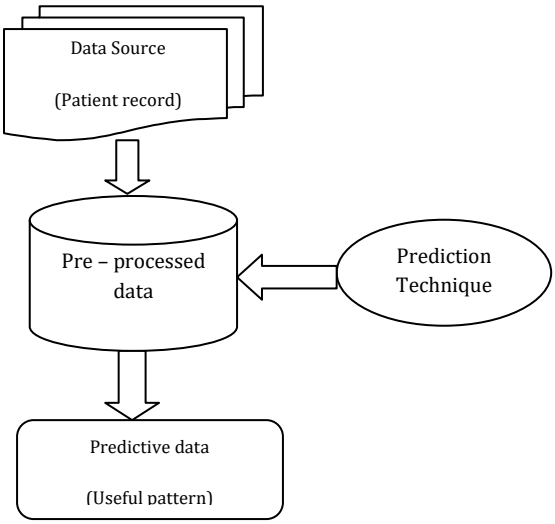

Figure 3. Healthcare Prediction

IV. PREDICTION TECHNIQUES

1. Artificial Neural Networks (ANN)

An artificial neural network is a mathematical model based on biological neural networks. It consists of an interrelated group of artificial neurons and processes information using a connectionist approach to computation. Neurons are structured into layers. The input layer consists of the original data, while the output layer nodes represent the classes. There may be several hidden layers. A main feature of neural networks is an iterative learning process in which data samples are presented to the network one at a time, and the weights are adjusted to predict the correct class label. Advantages of neural networks include their high tolerance to noisy data, as well as their ability to classify patterns on which they have not been trained.

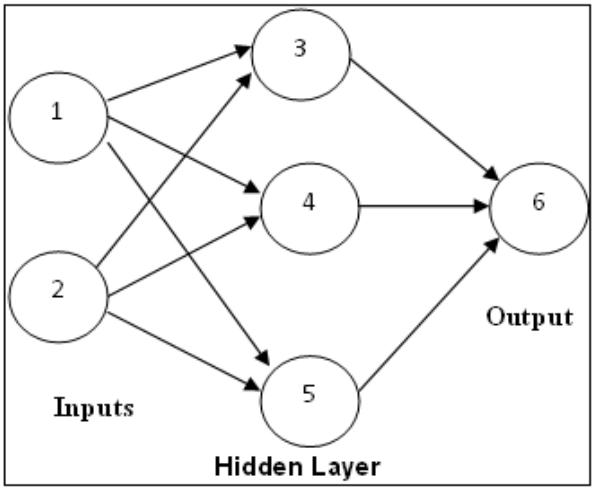

Figure 5. A neural network

A main concern of the training phase is to focus on the interior weights of the neural network, which is used according to the transactions used in the learning process. For each training transaction, the neural network receives in addition the expected output.

2. Bayesian Classifier

Bayesian classifier is a statistical classification approach based on the Bayes theorem.

Theorem:

To calculate probability of A given B, P (B given A) = P (A and B)/P (A) the algorithm counts the number of cases where A and B occurs simultaneously and divides it by the number of cases where A alone occurs. Let X be a data tuple, X is considered "Evidence", in Bayesian terms. Let H be some hypothesis, such that the data tuple X belongs to class C. P (H|X) is posterior probability, of H conditioned on X. P (H) is the prior probability of H in contract.

$$P(H|X) = \frac{P(H|X)P(H)}{P(X)}$$

$$Posterior = Likelihood * \frac{Prior}{Evidence}$$

3. Decision Tree

Decision tree uses the simple divide-and conquer algorithm. In these tree structures, leaves represent classes and branches signify conjunctions of features that lead to those classes. The attribute that most effectively splits samples into different classes is chosen, at each node of the tree. A path to a leaf from the root is found depending on the assessment of the predicate at each node that is visited, to predict the class label of an input.

Decision tree is fast and easy method since it does not require any domain information. In the decision tree inputs are divided into two or more groups continue the steps till to complete the tree as shown on Fig.4

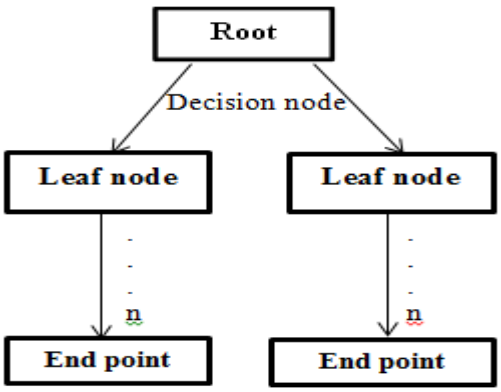

Figure 4. Decision tree Structure

Various decision tree algorithms as follows:

a) CART (Classification & Regression Tree)

b) C4.5 (Successor of ID3)

c) ID3 (Iterative Dichotomiser 3)

d) CHAID (CHI-squared Automatic Interaction Detector)

4. Support Vector Machine (SVM)

Normally SVM is the classification technique. Initially it developed for binary type classification later extended to multiple classifications. This SVM creates the hyper plane on the original inputs for effective separation of data points.

V. COMPARITIVE STUDY OF DIFFEEENT PREDICTION IN HEALTHCARE

A crucial of the data mining in the medical domain is better prediction through the practice and scientific observations. This section describes different data mining prediction applications which are in medical domain and work done by different researchers are given in detail. Different data mining tools are used to predict in different healthcare problems. The following list of medical issues has been studied and estimated in this section.

a. Eye disease
b. Cancer
c. Heart disease
d. Diabetics

There may be huge number of data mining techniques and data mining tools are available for predicting heart disease, various cancers, diabetics, eye disease and dermatological conditions. The following table presents comparison of disease, data mining techniques and the accuracy of the data mining techniques.

Table 1: Comparison of data mining techniques

S.NO	DISEASE	DATAMINING TECHNIQUE	ALGORITHM	ACCURACY (%)
1.	Diabetics	Decision tree(SPSS)	Chi-Square	75.21
2.	Diabetics	Decision tree	C4.5	91
3.	Diabetics	SVM	SMO	94.3
4.	Eye disease	Decision tree/Neural Network	Back propagation	92
5.	Urinary system disease	Neural network	Learning algorithm	99
6.	Lung cancer	Classification	Naïve Bayes	84.14
7.	Breast cancer	Decision tree(WEKA)	C4.5	86.7
8.	Parkinson's disease	Regression tree	-	93.75
9.	Heart disease	Classification	Navie Bayes	88.76
10.	Heart disease	Classification	Laplace Smoothing	86
11.	Heart disease	Classification	K-nearest neighbour's algorithm	98.24
12.	Heart disease	Classification	Navie Bayes	97.42
13.	Heart disease	Decision tree/IHDPS	-	89
14.	Heart disease	Classification	Navie Bayes	52.33
15.	Breast cancer	Classification	Random tree	100
16.	Breast cancer	Classification	Sequential Minimal Optimization (SMO)	96.19
17.	Breast cancer	Classification	SVM	95.7
18.	Breast cancer	Decision Tree	J48	94.5

VI. CONCLUSION

This paper is presented to study about the various data mining application in the healthcare sector to discover new range of pattern information. There is variety of data mining tools and techniques are available for health care diagnosis systems that are clearly defined. This data mining based prediction system reduces the human effects and cost effective one.

REFERENCES

[1] Muhamad Hariz Muhamad Adnan,Wahidah Husain, Nur'Aini Abdul Rashid(2012), "Data Mining for Medical Systems: A Review" ,International conferences on advances in computer and information technology.

[2] V. Krishnaiah et al,(2013)" Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques", International Journal of Computer Science and Information Technologies, Vol. 4 (1) ,39 – 45.

[3] Abdelghani Bellaachia, Erhan Guven," Predicting Breast Cancer Survivability Using Data Mining Techniques",Department of Computer Science, The George Washington University.

[4] Ravi Sanakal, Smt. T Jayakumari(2014)," Prognosis of Diabetes Using Data mining Approach-Fuzzy C Means Clustering and Support Vector Machine",International Journal of Computer Trends and Technology , vol. 11 (2).

[5] L. G. Kabari and E. O. Nwachukwu(2012)," Neural Networks and Decision Trees For Eye Diseases Diagnosis",INTECH.

[6] Qeethara Kadhim ,Al-Shayea and Itedal S. H. Bahia(2010),"Urinary System Diseases Diagnosis Using Artificial Neural Networks", IJCSNS International Journal of Computer Science and Network security, Vol.10 No.7.

[7] Dhanashree S.Medhekar,Mayur P.Bote,Shruti D.Deshmukh(2013),"Heart Disease Prediction using Naïve Bayes", International Journal Of Enhanced Research In Science Technology & Engineering ,Vol.2 Issue 3.

[8] Ms.Rupali R.Patil,(2014) "Heart disease prediction system using Naïve Bayes and Jelinek-mercer smoothing", International Journal Advanced Research in Computer and Communication Engineering, Vol.3, Issue 5.

[9] A.H. Hadjahmadi, and Taiebeh J. Askari(2012)" A Decision Support System for Parkinson's Disease Diagnosis using Classification and Regression Tree", The Journal of Mathematics and Computer Science Vol.4(2),.257 – 263.

[10] Hian Chye Koh and Gerald Tan." Data Mining Applications in Healthcare",Research Gate.

[11]M. Duraira and V. Ranjani(2013)," Data Mining Applications In Healthcare Sector: A Study", International Journal Of Scientific & Technology Research Vol. 2, Issue 10.

[12] Hlaudi Daniel Masethe and Mosima Anna Masethe(2014)," Prediction of Heart Disease using Classification Algorithms", Proceedings of the World Congress on Engineering and Computer Science Vol II , 22-24.

[13] Jyoti Soni, Ujma Ansari and Dipesh Sharma(2011)," Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", International Journal of Computer Applications ,Vol. 17– No.8.

[14] R. Chitra and V. Seenivasagam(2013)," Review Of Heart Disease Prediction System Using Data Mining And Hybrid Intelligent Techniques", ICTACT journal on soft computing, Vol. 03, Issue 04.

[15] S. Syed Shajahaan, S. Shanthi and V. ManoChitra(2013)," Application of Data Mining Techniques to Model Breast Cancer Data", International Journal of Emerging Technology and Advanced Engineering, Vol. 3, Issue 11.

[16] Vikas Chaurasia and Saurabh Pal(2014)," A Novel Approach for Breast Cancer Detection using Data Mining Techniques",International Journal of Innovative Research in Computer and Communication Engineering, Vol. 2, Issue 1.

[17] Ahmad LG et al.,(2013)," Machine Learning Techniques for Predicting Breast Cancer Recurrence", Health & Medical Informatics, Health Med Inform 2013.

[18] Ronak Sumbaly, N. Vishnusri and S. Jeyalatha(2014)," Diagnosis of Breast Cancer using Decision Tree Data Mining Technique", International Journal of Computer Applications, Vol. 98– No.10.

[19] K. Rajalakshmi & Dr. S. S. Dhenakaran(2015)," Analysis of Datamining Prediction Techniques in Healthcare Management System", International Journal of Advanced Research in Computer Science and Software Engineering,Vol.5,Issue4.