RESEARCH ARTICLE OPEN ACCESS

Data Mining Techniques in Healthcare: A Case Study

Chirag [1], Komal Sharma [2]
Assistant Professor (CSE) [1], RPSDC, Mahendergarh
Cognizant Technology Solutions [2], Gurgaon
Haryana - India

ABSTRACT

Data mining is the process of discovering information through large set of database and transform it into a understandable structure for further use, it can help researchers to gain both narrative and deep insights of exceptional understanding of large biomedical datasets. Manually analysing, classifying, and summarizing the data is impossible because of the unbelievable increase in data. Data mining can display new biomedical and healthcare knowledge for clinical decision making. Medical analysis is very important but complex problem that should be performed fairly and accurately. The goal of this paper is to discuss the techniques of data mining to solve the complex problem of prediction in Medical diagnosis with their advantages and disadvantages.

Keywords **:-** Data Mining, Decision Tree, K means Clustering, Naïve Bayes, and KDD Process.

I. INTRODUCTION

Data Mining is the process of getting useful information in the large database or you can say Data mining is the non-trivial process of knowing valid, novel, potentially useful, the ultimately understandable information from data. Rapid advance in data collection and storage technology have enabled organizations to accumulate vast amount of data. However extracting useful information has proven extremely challenging, so data mining is the technique that blends traditional data analysis methods with sophisticated algorithms for processing large amount of data for extracting meaningful information from that vast data. Data mining also provides capabilities to predict the outcome of a future observations, such as predicting whether a newly arrived patient have what kind of disease in past. Data Mining is also revealed as necessary process where best methods are used to extract the data patterns by passing through miscellaneous data mining processes.

Data mining is an integral part of discovering knowledge in large databases (KDD), which is process of converting vast data into useful or meaningful information, as shown in figure 1. This process is consists of a series of transformations steps, from data processing to post processing of data mining results.

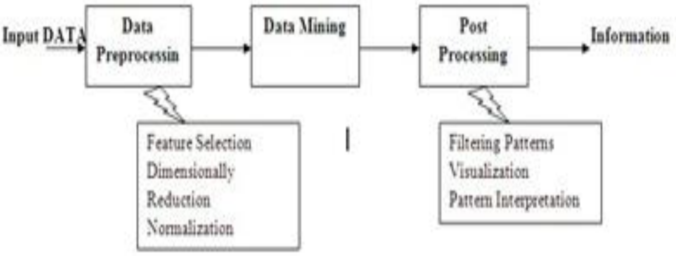

Figure 1 Process of knowledge discovery in database (KDD)

Here, Input data can be stored in many formats (flat files, spreadsheets, or relational tables form). Pre-processing means transform this input data (raw data) into appropriate format for analysis. Post processing is step that ensures that only valid and useful results are incorporated into decision support system.

II. DATA MINING TECHNIQUES IN MEDICAL SCIENTIFIC DISCIPLINE

In medical data mining, some techniques like Multilayer Perception, Neural Network, J48 Decision Tree, KNN, K-Clustering, Random Forest has been widely used. We have briefly discussed these techniques with their pros and cons:

A. Neural Network

The term 'Neural' is derived from the human (animal) nervous system's basic functional unit 'neuron' or nerve cells which are present in the brain and other parts of the human (animal) body.

The inventor of the first neurocomputer, Dr. Robert Hecht-Nielsen, defines a neural network as − "Neural Network is a computing system made up of a number of simple, highly interconnected processing elements, which process information by their dynamic state response to external inputs."[1]. the first artificial neuron was produced in 1943 by the neurophysiologist Warren McCulloch and the logician Walter Pits. Neural network provide powerful mechanism to help the physicians to review, model and make sense of complex clinical data across medical applications. [2-3]

1) ***Why use Neural Network:*** Neural Networks are the biologically inspired simulations performed on the computer to perform certain specific tasks like clustering, classification, pattern recognition etc., It can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques [4].

2) ***Basic Structure of Neural Network:*** The human brain is composed of 86 billion nerve cells

called **neurons.** They are connected to other thousand cells by **Axons** [5] **(**When the sum reaches a threshold value, neuron fires and the signal travels down the axon to the other neurons.**)**. Stimuli from external environment or inputs from sensory organs are accepted by **dendrites** (It receives signals from other neurons). These inputs create electric impulses, which quickly travel through the neural network. A neuron can then send the message to other neuron to handle the issue or does [6] not send it forward [5].

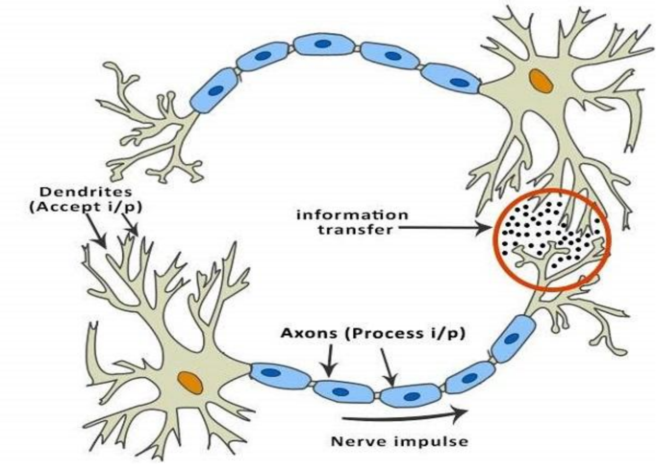

Figure 2 Human Brain Structure (Link For Image :- (**https://www.tutorialspoint.com/artificial_intelligence/artificial_intelligence_neural_networks.htm**)

3) ***A simple Neuron:*** An artificial neuron is a device with many inputs and one output. The neuron has two modes of operation; the training mode and the using mode. In the training mode, the neuron can be trained to fire (or not), for particular input patterns. In the using mode, when a taught input pattern is detected at the input, its associated output becomes the current output. If the input pattern does not belong in the taught list of input patterns, the firing rule is used to determine whether to fire or not.

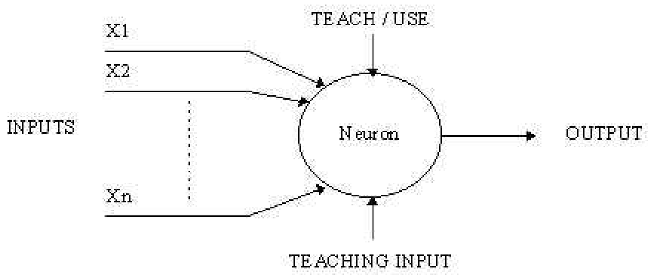

Figure 3 A Simple Neuron

4) ***Advantage of Neural Network:***

1. Error prone
2. Robust in noise environment
3. Neural Network (NN) is very efficient

5) ***Disadvantage of Neural Network:***

1. NN is high complexity model with long duration of training
2. Local minima
3. NN is over fitting

B. *Decision Tree*

There are some standard tree algorithms which are implemented: ID3 and C4.5 (called version J48). J48 is mostly used as compared to ID3 as J48 yields better results in any context [7]. In the WEKA data mining tool, J48 is an enhanced version of C4.5 algorithm. The decision tree generated by C4.5 is used for classification process. [8-9]

1) ***What is a Decision Tree?:*** A decision tree is a classification scheme which generates a tree and a set of rules, representing the models of different classes, from a given data set or you can say tree-shaped diagram used to determine a course of action or show a statistical probability. Each branch of the decision tree represents a possible decision, occurrence or reaction.

Decision Tree has three types of nodes in it:-

- **A Root Node** is the first node of any tree which has no incoming edges and zero or more outgoing edges.
- **Internal Node** which has exactly one incoming edge and two or more outgoing edges.
- **Leaf** which has exactly one incoming edge and no outgoing edges. A leaf is also known as Terminal Node. [10]

The set of records available for developing classification method id generally divided into two disjoint subsets – a training set and a test set. In order to have a clear idea of decision tree, let us consider the following data set- the training data set (See table 1).

Table 1 Training Data Set

OUTLOOK	TEMP (^{0}F)	HUMIDITY (%)	WINDY	CLASS
Sunny	79	90	true	No play
Sunny	56	70	false	play
Sunny	79	75	true	play
Sunny	60	90	true	No play
Overcast	88	88	false	No play
Overcast	63	75	true	play
Overcast	88	95	false	play
Rain	78	60	false	play
Rain	66	70	false	No play
Rain	68	60	true	No play

Rule 1 if it is sunny and the humidity is not above 75% then play

Rule 2 if it is sunny and the humidity is above 75% then don't play

Rule 3 if it is overcast then play

Rule 4 if it is rainy and don't windy then play

Rule 5 if it is rainy and windy then don't play

Decision Tree with the help of training data set

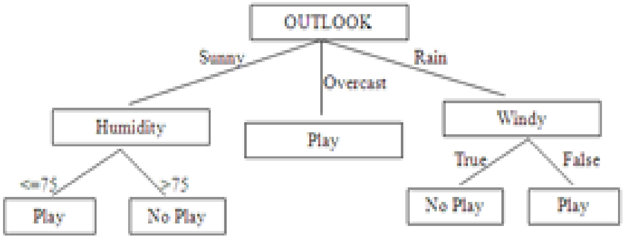

Figure 6 A Decision Tree

2) ***Advantage of Decision Tree:***

1. Decision Trees are able to generate understandable rules.
2. They are able to handle both numerical and categorical attribute.
3. Provide clear indication that which fields are most important for prediction.

3) Disadvantage of Decision Tree

1. Some decision trees can only with binary valued target classes.
2. The process of making a decision tree is computationally expensive.

C. K-Nearest Neighbor

K-Nearest Neighbor is a very simple and most powerful statistical unsupervised clustering approach. KNN can also work with least information or no prior information of data distribution. KNN also known by various names: Lazy learning, memory based reasoning, example based reasoning, instance based learning, case based reasoning.

Example: - Rote Classifier

The justification for using nearest neighbors is best exemplified by following saying: “If it walks like a duck, quacks like a duck, and looks like duck, then it’s probably a duck.”[11]

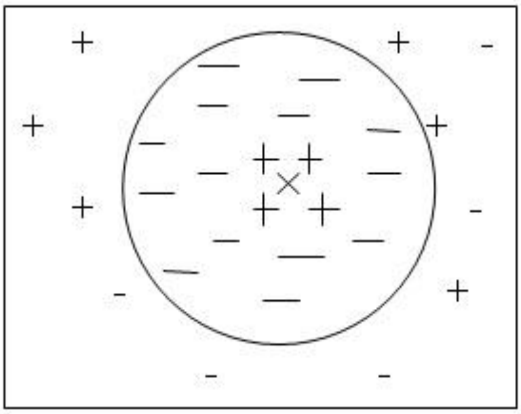

Figure 7 k-nearest neighbor classification

1) ***Algorithm:*** The K-nearest neighbor classification algorithm is as follow:

```
1 Let k be the number of nearest neighbor and D is the set of training examples.
2 for each test example z=(x', y') do
3     Compute d (x', x), the distance between z and every example, (x, y) € D.
4 Select D_z subset of D, the set of k closest training example to z.
5 Y' = argmax $\sum_{(xi,yi)€ Dz} I(v = yi)$
6 end for
```

2) ***Advantage of KNN:***

1. KNN is easy to use and understand.
2. In KNN fast training required.
3. KNN has high presentation ability.
4. In KNN there is no optimization.

3) ***Disadvantage of KNN:***

1. In KNN computation speed is slow.
2. KNN is sensitive to representation.
3. KNN has slow testing.
4. KNN requires huge area for storage.

D. K-means Clusterin

It is a simplest, quantitative and iterative method used for aggregating enormous sets of data.. An algorithm used to organize the objects based on features into number of clusters. Its main purpose is to define k centers, one for every cluster. These centers should be placed by a deceptive means as different location needs different results. [3]

1) ***K-means clustering for precise data:*** The classical K-means clustering algorithm which aims at finding a set C of K clusters C_j with cluster mean c_j to minimize the sum of squared errors (SSE). The SSE is usually calculated as follows:

$$\sum_{j=1}^{K}\sum_{x_i\in C_j}\|\mathbf{c}_j-\mathbf{x}_i\|^2 \qquad \ldots\ldots (1)$$

Where ‖ . ‖ is a distance metric between a data point $\mathbf{x}_i$ and a cluster means $\mathbf{c}_j$. For example, the Euclidean distance is defined as:

$$\|\mathbf{x}-\mathbf{y}\|=\sqrt{\sum_{i=1}^{V}|x_i-y_i|^2} \qquad \ldots.. (2)$$

The mean (centroid) of a cluster C_i is defined by the following vector:

$$\mathbf{c}_j=\frac{1}{|C_j|}\sum_{i\in C_j}\mathbf{x}_i \qquad \ldots\ldots (3)$$

The K-means algorithm is as follow

Assign initial values for cluster means c_1 to c_k

```
1 repeat
2  for i=1 to n do
3      Assign each data point $x_i$ to cluster $C_j$ where ‖ $c_j$-$x_i$‖ is the minimum
4  end for
5  for j=1 to K do
6      Recalculate cluster mean $c_j$ of cluster $C_j$
7  end for
8 until convergence
9 return C
```

2) ***K-means clustering for uncertain data**:* In order to take into account data uncertainty in the clustering process, Chau, et.al [10] propose a clustering algorithm with the goal of minimizing the expected sum of squared errors E (SSE). Notice that a data object $\mathbf{x}_i$ is specified by an uncertainty region with an uncertainty f($\mathbf{x}_i$). Given a set of clusters, C_j's the expected SSE can be calculated as:

$$\begin{aligned} &E\left(\sum_{j=1}^{k}\sum_{i\in C_j}\left\|\mathbf{c}_j-\mathbf{x}_i\right\|^2\right)\\ &=\sum_{j=1}^{k}\sum_{i\in C_j}E\left(\left\|\mathbf{c}_j-\mathbf{x}_i\right\|^2\right)\\ &=\sum_{j=1}^{k}\sum_{i\in C_j}\int\left\|\mathbf{c}_j-\mathbf{x}_i\right\|^2 f(\mathbf{x}_i)d\mathbf{x}_i \end{aligned} \quad ......\ (4)$$

Cluster means are then given by:

$$\begin{aligned} \mathbf{c}_j &= E\left(\frac{1}{|C_j|}\sum_{i\in C_j}\mathbf{x}_i\right)\\ &=\frac{1}{|C_j|}\sum_{i\in C_j}E(\mathbf{x}_i)\\ &=\frac{1}{|C_j|}\sum_{i\in C_j}\int\mathbf{x}_i f(\mathbf{x}_i)d\mathbf{x}_i \end{aligned} \quad ....\ (5)$$

They also propose a new K-means algorithm for clustering uncertain data.

Assign initial values for cluster means c_1 to c_k

```
repeat
  for i=1 to n do
    Assign each data point $x_i$ to cluster $C_j$ where
    $\| c_j - x_i \|$ is the minimum
  end for
  for j=1 to K do
    Recalculate cluster mean $c_j$ of cluster $C_j$
  end for
until convergence
return C
```

3) ***Advantage of K-Means Clustering:***

1. Fast
2. Easily Understand
3. Fairly Efficient

4) ***Disadvantage of K-Means Clustering:***

1. Algorithm Fails for nonlinear dataset
2. Unable to handle turbulent data and outliers

III. DISCUSSION OF PAPERS

In this literature Survey we present an overview of the current research being carried out using the data mining techniques for the diagnosis and prognosis of various diseases. The following algorithms we discussed: Decision Trees, Artificial neural network. [7]

The Table 2 below shows the survey made in various techniques algorithms used in Medical field for diagnosis.

Table 2: Survey representing the use of various techniques algorithms in medical field

Author	**Year**	**Knowledge Type**	**Knowledgeable Outcomes**	**DM Techniques /Applications**
Jyoti Soni et.al [12]	2011	Heart Disease Prediction	Decision tree outperforms and sometimes Bayesian classification's having similar accuracy as Decision tree.	Classification: Clustering Bayesian classification, Neural Networks Decision Tree, KNN
Samar Al Qarzaie et.al [13]	2011	Breast Cancer Disease	WEKA tool is used to give 93.4675% accuracy in testing set and in the training set it yields 96.8% accuracy	Classification: Decision Tree
Arvind Sharma et.al [14]	2012	Blood Donors	By using WEKA tool, J48 decision tree acquires 89.99% accuracy	Classification: J48 Decision Tree
Dr. Bushra M. Hussan [15]	2012	Prediction of medical data by K means Clustering	On changing the instances it shows 97% of accuracy.	Classification: K-means, Clustering
Shweta Kharya [16]	2012	breast cancer diagnosis and prognosis	Decision tree is a best predictor with 93.62% accuracy	Classification: Neural Network, Association, Naive. Bayes, C4.5 decision tree algorithm
M. Durairaj et.al [17]	2013	Applications in healthcare sector	WEKA tool gives 97.77% accuracy for cancer prediction and about 70% accuracy for success rate of IVF treatment.	Classification: Artificial Neural Network

Aarti Sharma et.al [18]	2014	Applications of Data Mining	Decision tree shows good accuracy for processing raw data into information and find patterns	Classification: association, rule clustering, prediction and Evaluation pattern
Mehak Naib et.al [19]	2014	Predict Primary Tumors using multiclass classifier approach	Multiclass classifier gives good accuracy by using WEKA tool than binary classifier	Classification: Random Forest

IV. CONCLUSION AND FUTURE SCOPE

The motive of this paper was to discover the most typical data mining algorithms. The ideas of future work consider the evaluation of chosen algorithms on the basis of chosen medical dataset. Different algorithms can be applied on built-in dataset and the algorithm which gives best result will be applied on the test dataset. The experiments would be conducted on the selected medical records which design the analysis even more accurate. One more idea is taking other algorithms also and compares their performance in medical field. And these types we evolve a new assist in scheming Medical Decision Support Systems by selecting one of the most acceptable algorithms. Other methods which are not mentioned in this paper on comparison basis and can discover the best one by assessing the advantages and limitations of the prevailing one.

ACKNOWLEDGEMENT

We express our sincere gratitude towards the management of RPS Degree College and Cognizant Technology Solutions, Gurgaon for providing us financial assistance for the research work. We acknowledge our gratitude to Dr. Yashpal Sharma and Dr. D.R. Bhardwaj for sharing their pearls of wisdom with us during the course of this research.

REFERENCES

[1] Mr. Milind S. Khairnar., Use of Neural Networks for Developing Clinical Decision Support System - A Review, International Conference on Recent Trends in Computer and Electronics Engineering (ICRTCEE 17) (2017), Vol. No.- 5, (2321-8169).

[2] Mittal P., Study and Analysis of Predictive Data Mining Approaches for Clinical Dataset, International Journal of Computer Applications (2013), and Vol. no. - 63, No. 3.

[3] Rani K. Analysis of Heart Dksiseases Dataset Using Neural Network Approach, International Journal of Data Mining and Knowledge Management Process (IJDKP) (2011), Vol. 1, No. 5.

[4] Sumit Singh, Varun Vashishtha, Tapodhan Singla., Artificial Neural Network, International Journal of Research (IJR) (2014), Vol-1, (2348-6848).

[5] Artificial Intelligence, Retrieved from www.tutorialspoint.com/artificial_intelligence/, 2017

[6] M.K.Karthik, A. Vasudevan, K.K Anuraja, K. manojkumar., Failure load prediction of composite laminates using acoustic emission, International journal of engineering sciences & research technology (2017) (2277-9655).

[7] Chink Gera, Kirti Joshi., A Survey on Data Mining Techniques in the Medicative Field, International Journal of Computer Applications (2015) (0975 – 8887), Volume 113 – No. 13

[8] Upadhyay N., A Survey on the Classification Techniques in Educational Data Mining, International Journal of Computer Applications Technology and Research (2014), Vol. No.- 3,(2319-8656).

[9] Chaurasia V., Mining Approaches to Detect Heart Diseases, International Journal of Advanced Computer Science and Information Technology (2011), Vol. 2, (2296-1739), 56-66.

[10] Michael Chau, Reynold Cheng, Ben Pao, Uncertain Data Mining: A New Research Direction, Proceedings of the Workshop on the Sciences of the Artificial (2005).

[11] Vipin kumar., Handbook of introduction to data mining, India, 2008.

[12] Soni J, Predictive Data Mining Diagnosis: An overview of Heart Diseases Prediction, International Journal of Computer Applications (0975-8887) (2011),Vol. 17.

[13] Qarzai S., Using the Data Mining Technique for Breast Cancer Early Predication (2011).

[14] Sharma A., Predicting the Number of Blood Donors through their Age and Blood Group using Data Mining Tool, International Journal of Communication and Computer Technologies (2012), (2278-9723), Vol. no.-6.

[15] Hussan D., Data Mining based Prediction of Medical data using K-means algorithm, Basrah Journal of Science(A) (2012), Vol.30(1).

[16] Kharya S., Using Data Mining Techniques for Diagnosis and Prognosis of Cancer Disease, International Journal of Computer Science, Engineering and Information Technology (2012), Vol. no.-2.

[17] Durairaj M., Data Mining Applications in Healthcare Sector: A Study, International Journal of Scientific and Technology Research (2013), Vol. no.-2, (2277-8616),

[18] Sharma. A, Applications of Data Mining: A survey Paper, International Journal of Computer Science and Information Technologies (2014), Vol. no.-5(2), (0975-9646).

[19] Naib M., Predicting Primary Tumors using Multiclass Classifier Approach of Data Mining, International Journal of Computer Application (2014), Vol. no-96.