International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 04 | Apr 2021 www.irjet.net p-ISSN: 2395-0072

Analysis of Data Mining Tools Used in Healthcare Domain

Jhumpa Sarma

Student, Dept. of Computer Science and Engineering, Jorhat Engineering College, Jorhat, Assam, India

---

Abstract – *Adaptation of information technology has led to creation of several applications in healthcare informatics. Data Mining has gained popularity in different research areas due to its numerous applications and methodologies used to mine the information in the correct manner. It is the process of analyzing, extracting data and furnishing the data as knowledge which forms the relationship within the available data. In the last decade, there has been an increase in usage of data mining techniques especially on medical data for determining useful patterns and trends which are used in analysis and decision making. Data mining has an infinite potential to utilize healthcare data effectually to predict any disease. As widely said "Prevention is better than cure," prediction of diseases and epidemic outbreak would lead to an early prevention of occurrence of a disease. This paper features data mining techniques like classification, clustering and highlights related work for analysis and prediction of human disease.*

Key Words: Data Mining, Classification, Association, Healthcare.

1. INTRODUCTION

Significant advances in information technology results in excessive growth of data in healthcare informatics. Healthcare informatics data may include hospital details, patient's details, disease details and treatment cost. These huge data which are generated from different sources and format can have irrelevant attributes and missing data. Application of data mining techniques is a key approach to extract knowledge from large disease data. Data mining processes include framing a hypothesis, gathering data, performing pre-processing, estimating the model, understanding the model and drawing suitable conclusions [2] Let us understand the types of algorithms that exists in data mining and understand how they are functioning.

Data mining came into existence in the middle of 1990s. In common, Knowledge Discovery (KDD) and Data Mining are used interchangeably but many researchers assume that both terms are dissimilar as Data Mining is one of the most vital stages of the KDD process. Knowledge discovery in databases is the process of finding useful information and patterns in data. It uses algorithms to extract the information.

According to Fayyad et al., the Knowledge Discovery in database is systematized in various stages. The first stage is selection of data in which data is gathered from different sources, the second step is the pre-processing the selected data, the third stage is transforming the data into suitable format so that it can be processed further, the fourth stage consist of Data Mining where suitable Data Mining techniques are applied on the transformed data for extracting valuable information. The last stage is evaluation. Various stages of knowledge discovery in databases process

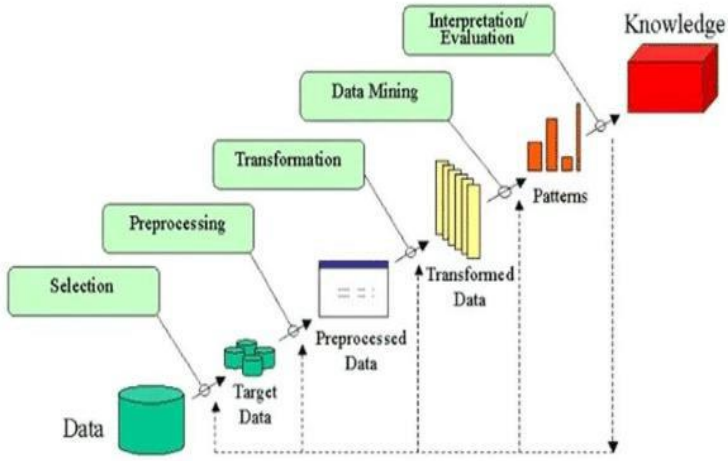

Fig1. Stages of Knowledge Discovery Process

are shown in Fig1.

Selection step involves the collection of heterogenous data from various sources for processing. Medical data used in real life may be incomplete, inconsistent, noisy and irrelevant which requires a selection process that gathers the important data from which knowledge is to be extracted.

Pre-processing step involves performing basic operations of trying to find the missing data, eliminating the noisy data, detecting or removing outliers and resolving inconsistencies among the data.

Transformation step transforms the data into forms which are suitable for mining by performing aggregation, smoothing, normalization, generalization and discretization.

Data Mining step involves choosing the data mining algorithms and using the algorithms to generate previously unknown and hypothetically beneficial information from the data which is stored in the database.

Evaluation stage is the presentation of mined patterns in understandable form. In this step, the mined patterns are interpreted and evaluation of the outcomes are prepared using statistical justification and significance testing.

2. DATA MINING TECHNIQUES

Data Mining Algorithms are classified in two categories: descriptive model (or unsupervised learning) and predictive model (or supervised learning). The purpose of Descriptive data mining model is to discover patterns in data and identifying the association between attributes represented by data. On the other hand, the purpose of Predictive data mining model is largely to predict the future outcome than the existing behavior. [3]

In order to increase the capability for making appropriate conclusions regarding patient health from raw facts and figures, Data mining techniques like association, classification and clustering are used.

2.1 Clustering

Clustering is a data mining technique which makes cluster of objects that have similar characteristics using automatic techniques. Clustering does not have predefined classes. Clustering algorithms discover collections of the data such that objects in the same cluster are more identical to each other than the other groups. [4] Different types of Cluster techniques include K-means, Fuzzy C-means (FCM), Rough C-means (RCM), Rough Fuzzy – C means (RFCM), Robust RFCM (rRFCM), hierarchical and Gaussian mixture.

2.2 Classification

Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model which can classify population records at large. This approach frequently makes the use of decision trees or neural network-based classification algorithms. The data classification process involves Learning and Classification. In Learning, the training data is analysed by classification algorithms. In classification, test data is used to estimate the accuracy of the classification rules.

2.2 Association

Association rule mining is a well-researched method for finding interesting relations between different data in large databases. It serves the purpose to identify well-built rules discovered in databases using different procedures of importance based on input dataset. An integrated approach of using classification and association have improved the capabilities of Data Mining. Soni et al., has used this integrated approach for studying healthcare data. This integrated approach is useful for determining rules in the database and with the help of these rules, an effective classifier can be raised. The study made an experiment on the data of heart patients and generated rules by a weighted associative classifier. [7]. Hence, Association has tremendous influence in healthcare domain for identifying relationships among different diseases, symptoms of the disease and the state of human health.

Table -1: Data Mining Tasks and Techniques

Data Mining Tasks and Techniques	
Data Mining Task	Data Mining Algorithm and Technique
Classification	Neural Network Support Vector Machines, Decision Trees, Genetic Algorithms, Rule Induction
Clustering	K-means FCM(Fuzzy C-means),RCM(Rough C-means),Rough Fuzzy-Cmeans (RFCM)
Association and Link Analysis	Association Rule Mining

3. APPLICATION OD DATA MINING TECHNIQUES IN HEALTHCARE

In aspects of prediction and decision making, the use of Data mining techniques have a broad expansion in healthcare industry with respect to various diseases like heart diseases, liver diseases, cancer, pneumonia, diabetes and others.

Table 2 provides the summary of medical data classification regarding the resolved difficulties that are solved, convenience in medical data mining or implementation of the tools.

Table -2: Summary of Medical Data Classification and the resolved difficulties which are solved

Summary of Medical Data Classification and the Resolved Diffulties			
Author(s), Year	Medical Dataset	Technique(s)	Comments
Khanmohammadi & Chou, 2016	Six datasets (UCI Repository)	GMBD	Discretization process was more concise for representation of continuous variable
Aswal & Ahuja 2016	six datasets (UCI Repository)	K-NN, SVM	Classification of bio medical data
Long, 2015	heart disease	Rough sets based attribute reduction and interval type-2 fuzzy logic system (IT2FLS)	heart disease diagnostic system using rough sets based on attribute reduction and IT2FLS
Zuo et al. 2013	Parkinson Disease	Fuzzy-KNN approach	Familiarised an adaptive Fuzzy K-NN approach for diagnosing the disease
Ghofrani et al.	X-Ray	K-NN & SVM	Slow testing,

2014	dataset		scale dependable.
Polat et al. 2007	Breast Cancer and Liver Disorders dataset	Fuzzy-AIRS	Modelling and analysis of medical data
Tang et al. 2009	Diabetes, Cancer	k-Nearest Neighbour	Classification of Disease using K-NN
Xing etal. 2007	coronary heart disease	Decision Tree Algorithms such as C4.5, C5, and CART	Prediction models

Acronym:

Genetic k-Nearest Neighbour (GkNN), Support Vector Machine(SVM), k-nearest neighbour(K-NN), Gaussian mixture model based discretization (GMBD), Bayesian Ying Yang (BYY).

The guidelines on the usage of different data mining techniques include: Identification of the unnecessary attributes which impedes the processing task is crucial before application of classification technique. Besides acting as noise, they also affect the classifier performance. The researchers have found out that there is no classifier that generates the best result. For checking the performance of classifiers, each dataset is divided into – training and testing. A classifier which is tested using a testing dataset, is chosen based on its performance. For both training and test data, Cross Validation method is conducted to ensure accuracy.

The different classification algorithms mentioned below in Fig-1 are used to predict or analyze different diseases. [2]

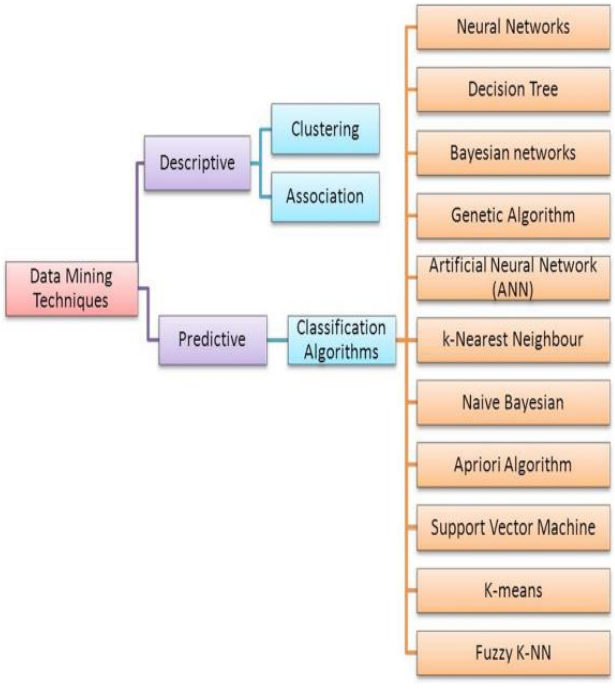

Fig -1: Different Techniques Used in Healthcare Domain

4. CONCLUSIONS

With the recent rise in the quantity of biomedical data that is gathered by electronic means, many researchers have started, or are eager to start, exploring these data. In this paper we have observed some data mining techniques that have been used for analyzing medical data. Classification is the major data mining technique which is primarily used in healthcare sectors for medical diagnosis and predicting diseases. This paper highlights the data mining techniques that are used for medical data for various diseases which are identified and diagnosed for human health.

REFERENCES

[1] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.
[2] Sheenal Patel & Hardik Patel, "SURVEY OF DATA MINING TECHNIQUES USED IN HEALTHCARE DOMAIN", International Journal Of Information Sciences and Techniques(IJIST), Vol.6, No.1/2, March 2016
[3] Pradnya P. Sondwale, "OVERVIEW OF PREDICTIVE AND DESCRIPTIVE DATA MINING TECHNIQUES", International Journal of Advanced Research in Computer Science and Software Engineering, Volume: 05 Issue: 04, April-2015.
[4] K.Sharmila & Dr.S.A.Vethamanickam, "SURVEY ON DATA MINING ALGORITHM AND ITS APPLICATION IN HEALTHCARE SECTOR USING HADOOP PLATFORM", International Journal of Emerging Technology and Advanced Engineering ISSN 2250-2459, Volume: 05, Issue: 01, January-2015.
[5] J.J.Tapia & E. Morett & E. E. Vallejo, "A CLUSTERING GENETIC ALGORITHM FOR GENOMIC DATA MINING", Foundations of Computational Intelligence, Studies in Computational Intelligence, Volume:204, 2009.
[6] L.Chang & C.H.Chen, "APPLYING DECISION TREE AND NEURAL NETWORK TO INCREASE QUALITY OF DERMATOLOGIC DIAGNOSIS", Expert Systems with Applications- Elsevier, Volume: 36, pp. 4035-4041, 2009.
[7] S. Soni & O. P. Vyas, "USING ASSOCIATIVE CLASSIFIERS FOR PREDICTIVE ANALYSIS IN HEALTH CARE DATA MINING", International Journal of Computer Applications, Volume: 04, No: 05, July-2010
[8] W.L.Zuoa & Z.Y.Wanga & T.Liua & H.L.Chenc, "EFFECTIVE DETECTION OF PARKINSON'S DISEASE USING AN ADAPTIVE FUZZY K-NEAREST NEIGHBOR APPROACH", Biomedical Signal Processing and Control, Elsevier, pp. 364373, 2013.
[9] O.Er & N. Yumusakc & F. Temurtas, "CHEST DISEASES DIAGNOSIS USING ARTIFICIAL NEURAL NETWORKS", Expert Systems with Applications- Elsevier, Volume: 37, pp. 76487655, 2010
[10] Usama Fayyad & Gregory Piatetsky & Padhraic Smyth, "Knowledge Discovery and Data Mining: Towards a Unifying Framework" KDD-96 Proceedings, 1996
[11] Yanwei Xing & Jie Wang & Zhihong Zhao & Yonghong Gao, "COMBINATION DATA MINING METHODS WITH NEW MEDICAL DATA TO PREDICTING OUTCOME OF CORONARY HEART DISEASE", International Conference on Convergence Information Technology, 2007