Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 08, August-2020

A Review on Data Mining Techniques and Challenges in Medical Field

R. Pallavi Reddy
CSE, GNITS
Hyderabad, India

Ch. Mandakini
CSE, GNITS
Hyderabad, India

Ch. Radhika
CSE, GNITS
Hyderabad, India

Abstract:- **The healthcare industry has witnessed an enormous evolution in producing huge amounts of medical data that have given rise to research in multiple areas. Many researchers reviewed and surveyed the healthcare, which is an active interdisciplinary field of data mining. Technological advances in information on health care, digitizing health records, have resulted in rapid growth of the healthcare sector. Electronic Health Record Systems (EHRs) are the data repositories which are the digitized format for the medical data storage. Healthcare sector manages enormous amounts of data that needs to be analyzed to provide a better solution for better decision making. The main challenge is how to use the data mining techniques to effectively discover useful and important information among the massive amount of data available. It plays a major role in the advancement and development of new techniques that work effectively for the huge data in healthcare. The related information is collected that demonstrates the importance of data mining in health care. This paper mainly focuses on the necessity of data mining in medical field, its applications in health sector, different predictive and descriptive data mining techniques that can be used in various applications of healthcare sector and challenges that are involved in mining the health data.**

Keywords – Healthcare; medical data; data mining; Electronic Health Record Systems (EHRs)

I. INTRODUCTION

Data mining is the process of evaluating the databases to extract new insights from them. Data mining is becoming more popular in healthcare now, a days. It offers great potential to the healthcare industry for enabling health systems to systematically use data and analytics to identify inefficiencies and best practices that improve care and reduce costs. Due to the exponential increase in the number of electronic health records, data mining holds incredible potential for health care services. Doctors and physicians previously hold patient information in the physical documents which was quite difficult. The digitalization and invention of new technologies eliminates human effort and makes data easy to analyze.

Data mining reshapes many industries, including the healthcare sector. Applications for data mining can unbelievably benefit all people involved in the healthcare sector. The data framework simplifies and automates the health-care organizations' workflow. The integration of data mining into data frameworks reduces the decision-making effort of health care institutions and provides new valuable medical knowledge. Electronic health records (EHRs) are common in health-care institutions. With improved access to vast volumes of patient data, healthcare professionals are now concentrating on maximizing the efficiency and consistency of using data mining in their organizations. Predictive models provide healthcare professionals with the best information support and knowledge. The goal of predictive data mining in medicine is to build an effective predictive model, provide reliable predictions, support physicians in improving their diagnosis and treatment planning process etc. Data mining may help clinicians decide the best courses of action, minimize instances of unknown medication reactions, enhance the quality and safety of patients, identify factors related to fraud in health insurance, match specialists to patient needs etc. Data mining helps the healthcare organizations to evaluate treatment effectiveness, saves patients' lives using predictive medicine, manage the customer relationship, to detect fraud and abuse and in many other applications.

II. DATA MINING PROCESS

The amount of data produced in the healthcare sector needs to be transformed into useful knowledge for decision making. Data mining is a great promise in healthcare which analyzes complexity of data to generate information. The data mining process helps to discover knowledge from the selection stage to knowledge discovery stage. The below figure 2.1explains the data mining process in the medical field.

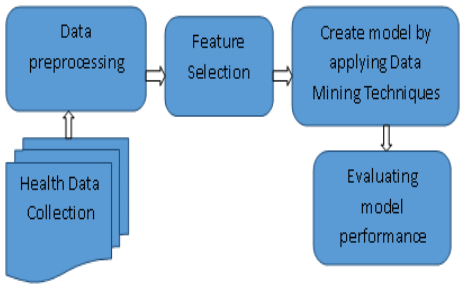

Fig 2.1: Data Mining Process in Medical Field

This section explains the data mining process for building a model and its performance assessment.

A. Data preprocessing:

It is a technique used in data mining to convert the raw data into a useful and efficient format. Data preprocessing steps

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 08, August-2020

include: Data Cleaning, Data Creation, and Data Reduction. The data can have several sections which are insignificant and missing. Software cleaning is done to handle data which is noisy, missing, etc. The data transformation is done to convert the data into appropriate form suitable for the mining process. This includes the generation of normalization, collection of attributes, discretization and generating hierarchy. Data mining is a technique used for managing enormous quantities of data. In these instances, analysis becomes harder when dealing with huge volume of data. To get rid of that, the technique of data reduction is used. It aims to increase the efficiency of storage and reduce the cost of data storage and analysis. It consists of data cube aggregation, collection of subset attributes, reduction of numerosity, and reduction of dimensionality.

B. Feature selection:

Feature Selection can be defined as selecting a minimum subset of features that are actually necessary for any data mining process. The feature set may be redundant and the efficiency may be reduced. Additionally, the feature selection minimizes the number of essential features needed to optimize model accuracy. It helps in reducing the space required by the feature set. This also eliminates the redundant noise that may be present in the feature set and thus improves the effectiveness of the algorithm for data mining. The objective of feature selection is to produce an efficient and cost-effective model.

Feature Selection consists mainly of four stages: subset development, subset evaluation, selection criterion and final sub-set feature. The feature set is checked in the first step after eliminating inconsistencies such as the null values and the redundancies. After searching for the feature set, the subset generation process starts. The attribute evaluator evaluates the generated subset. The subset generation and evaluation process continues until the selection criteria are met. The final subset feature set is selected only after completing the above process.

C. Creating a model:

The data mining model gets data from the mining structure and then analyzes the data using data mining algorithms. The mining structure stores information which defines the data source. A mining model stores data from the statistical process, such as patterns found as a result of the analysis. Each type of model creates different set of patterns, item sets, rules or formulas which can be used to makepredictions. The algorithms that can be used in model development process are decision tree, neural networks and logistic regression etc.

D. Evaluating model performance:

There are various approaches for evaluating performance of the model. Commonly used measurement criteria that are appropriate are: accuracy, sensitivity, specificity, precision, and F-measure. Accuracy is defined as the ratio of correctly classified cases. Sensitivity or recall measures the ratio of the actual positives that are correctly identified. Specificity measures the ratio of actual negatives correctly identified. Precision, also known as the positive predictive value (PPV), measures the ratio of true positives to predicted positive cases. F- Measure is the harmonic mean of both precision and recall.

III. TECHNIQUES USED IN DATA MINING:

The data mining techniques are of two types, descriptive and predictive. The descriptive analysis is used to mine the data and provide the latest information on past/ recent events. On the other hand, the predictive analysis provides answers to the future queries that move across using historical data as the chief principle for decisions.

The figure 3.1 explains the different data mining techniques that can be used in medical field.

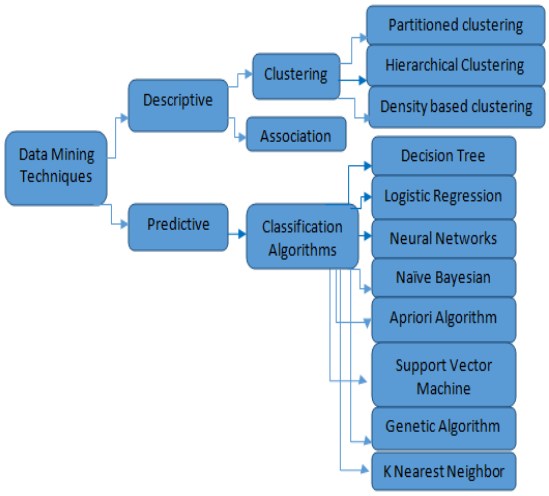

Figure 3.1: Data Mining Techniques

Classification:

It is a common technique of data mining and is used to categorize each item into one of a predefined set of classes or groups within a data set. Classification method utilizes a variety of mathematical techniques such as decision trees, neural networks, logistic regression, support vector machine, genetic algorithm, Bayesian networks. Classification software can learn from the dataset to predict future happenings. Data set features can be classified as low, moderate, high and very high in classification based on the symptoms of the diagnosed diseases. Classification is the most commonly used method of identification, prediction and optimization in the healthcare industry.

Decision Tree:

It is the widely used technique of data mining as users can easily understand its pattern. In this technique, a simple question or condition which has compound answers is the root of the decision tree. Then each answer leads to a group of questions or conditions that help to determine the data so that the ultimate decision can be based on it. It is a tree-like model of data in the database. Decision making is used to predict possible events and helps to increase the accuracy of the result. There are terminal and non-terminal nodes of a decision tree. Every non-terminal node on a data item

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 08, August-2020

represents a test or a condition. Decision trees classify the instances by sorting them down to the terminal nodes from the non-terminal. The selection of output branch depends entirely on the test result. Decision trees are commonly used in the analysis of operations research to calculate the conditional probabilities. With the support of decision trees, the best alternatives can be selected and, based on maximum knowledge gain, the traversal from root to leaf node indicates a unique class separation.

Neural Networks:
Neural networks are intelligent targets for data mining, as they are structured to function just like the human brain and aim to find hidden links between the data. An artificial neuron is a data processing unit which receives weighted input values from other attributes, transforms the value received according to some formula, and sends output to other attributes. Neural network is the best classification algorithm prior to the invention of decision trees and Support Vector Machine. The main purpose of using neural networks is to recognize patterns and perform the classification tasks. By adjusting the weights it helps to minimize the error due to its adaptive nature. These neurons work together to generate the output function in parallel. In the learning stage, the network must learn to predict the correct class label of the input by changing the weights. Neural networks have added advantage because unlike simple modeling methods they can predict nonlinear relationships. Neural networks are instrumental in analyzing medical data. Neural networks are used as one of the most popular algorithms for data processing in medicines. Neural network applications in this area include tissue classification, disease prediction, and drug production. Predicting cardiac disease can be done using a neural network. Multi-Layer Neural Network (MLNN) uses hidden layers with the aid of which it solves the problem of nonlinear sets classification. Usually, those hidden layers are interpreted as hyper-planes. Such neural networks are used to classify different data categories.

Logistic regression:
The label attribute is predicted based on the values of the input attributes. It describes the relation between the label attribute and the set of input attributes describing it. In the field of healthcare, logistic regression is used to predict diseases. It is primarily a statistical instrument used in data mining. It only analyzes the logistic and non-linear categorical data.

K-Nearest Neighbor (K-NN):
It uses techniques of classification and regression, and is an easy tool to use. In KNN, new data introduced into the database are analyzed by finding the subset of that data set to get the optimal solution for predicting an accurate outcome. This technique is used as a yardstick for predicting heart disease.

Support Vector Machine (SVM):
It is applied using classification and regression in a supervised learning to evaluate the data. It divides into two classes of hyper plane line. SVM will automate the processes making it more effective. It is applied significantly in healthcare for the identification of predictive features. The hyper- plane is the division between two outputs in a binary classification task, such as predicting ICU mortality. The key task of using hyper-planes is to maximize the separation between data points. For noisy data, error is minimized by maximizing the margin between two separate classes of examples and defining the hyper- plane as the center line of the separating space. Two forms of SVMs exist. The first is Linear SVMs, which separate the data points using a linear boundary for decision. It performs well on datasets, which can be easily split into two parts. The complex datasets are difficult to define using a linear kernel that uses the second form of SVMs, i.e. non-linear SVMs that separate the datasets using nonlinear decision boundaries. The SVM shows accuracy in problems of binary classification such as valve classification, heart beat etc.

Genetic Algorithm:
The genetic algorithm is a genetic and selection based search and optimization technique. Genetic algorithms are used primarily in neural sets that act as a guide for the learning process of data mining algorithms, rather than pattern finding. These are often used to formulate hypothesis about variables and dependencies among them in the form of association rules or some other formalism in data mining. In a genetic algorithm, there is a population composed of many individuals that evolve to a state where fitness is maximized under specific selection rules. A population of rules is initially created at random, with each rule representing a solution to the problem. Instead pairs of rules are selected as parents which are usually the strongest rules. A genetic algorithm consists essentially of three operators-selection, crossover, and mutation. In selection, a suitable string is chosen on the basis of fitness for the breeding of a new generation, then crossover blends these suitable good strings to produce better offspring, then mutation changes a string locally so that the genetic diversity is retained from one generation to another. For the termination of the algorithm the population is evaluated in each generation. If the termination criteria are not met, it is again operated by the three operators and then it is evaluated again.

Bayesian Network:
Bayesian network is a specific form of network which represents uncertain domain information. It belongs to the category of graphical probabilistic models (GMs). Nodes in the Bayesian network represent the variables, and specific edges represent probabilistic dependencies. For each variable, Bayesian network defines two types of knowledge. In medical science, the Bayesian classifier is based on probability theorem and can be used as the logical process for conducting medical diagnosis, especially in automated decision support systems.

Apriori Algorithm:
It is an algorithm for the learning of frequent item set mining and association rule over relational databases. This continues by defining the regular individual items in the database and expanding them to larger and larger item sets, as long as those

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 08, August-2020

item sets appear in the database sufficiently often. This solves problems from the most difficult to the least and groups the data that needs to be analyzed. Once all the problems are resolved the algorithm will stop. It shows the relation between two inputs to distinguish consistent and inconsistent inputs. Different types of apriori algorithms can be used such as Hash table, reduction of transactions, partitioning etc. in healthcare for predicting diseases.

Clustering:
This process uses the automated technique to construct an useful cluster of objects which have similar characteristics [2]. This defines the classes and places objects in each class. In a finite set it identifies the data categories and is used to predict any outcome. Kmeansand xmeans are the algorithms used in the clinical phase and for the evaluation of outcomes. Clustering can be classified as:

a. *Partitioned Clustering:*
It classifies related features of the dataset into different groups and analysts to know the clusters being generated. The partitioned clustering algorithm assists in predicting and diagnosing symptoms for specific diseases.

b. Hierarchical Clustering:
It clusters dataset features in the form of hierarchy used in healthcare predictions. It can break down the cycle from top to bottom or the methods from bottom to up. The clustering strategies are classified into Agglomerative and Divisive. The clustering result is often presented in a dendrogram form.

c. Density Based Clustering:
This clusters related points that are gathered in datasets, points closer to each other. It detects anomaly in points gapped within themselves. It is the most used clustering algorithm because of the performance.

Association:
Association is one of the best known techniques in data mining. In association, a pattern is learned in the related transaction, based on a correlation between items. Association seeks relationship from datasets by classifying the data into others to predict and to give better outcome. It is used where better accuracy is required. The two categories are mining classification and the association rule mining. In association, no attributes are required to discover the rule for an unsupervised learning.

IV. DATA MINING TOOLS USED IN HEALTHCARE:

Data mining tools help to analyze the volumes of complex data based on the data set attributes that users specify in determining trends of occurrences. The software can be used for diagnosis, prediction, and management of diseases to extract knowledge and make decisions. The choice of choosing suitable software to solve a specific problem is difficult because of the availability of various software tools. The most common data mining tools are:

a. WEKA (Waikato Environment for Knowledge Analysis):
WEKA is a software tool that is used in data mining processes. It is a software that is developed using Java programming and runs on different operating systems. WEKA compliments several processes related to data mining. The software may link directly to the data, or from the java code. It uses Graphical User Interface (GUI) to control the performance and features.

b. KEEL (Knowledge Extraction based on Evolutionary learning):
To extract the pattern from datasets, KEEL uses clustering, regression, and classification. It is an open source software, but source program may be hidden. The KEEL data mining tools can be used to perform complete analysis.

c. R:
It is an open source program used for computation and statistical analysis. R software is of great benefit to the world of research and development and to the health sector.

d. KNIME (Konstanz Information Miner):
It is an open source software that is used for analyzing and modeling data. The features of Machine Learning and Data Mining are supported with KNIME software. KNIME has been used in clinical studies, disease identification and evaluation. KNIME can create work processes which can be recorded in various formats.

e. RAPIDMINER:
It is used to analyze data which supports data mining processes in business, finance, banking, insurance, medical and education. It is an open source program which is used in numerous fields of human endeavors.

f. ORANGE:
Orange is an open source software. It is represented by front end and back end features. The front end uses visual programming, while python libraries are used in the back end. It was developed using ++ and Python programming. In science, Orange is used for testing the genetics and the medical field using various algorithms and techniques that can be further used in the education field.

V. CHALLENGES

The major limitation of data mining in healthcare is the heterogeneous and voluminous pertinent raw data. This is correlated with data from different sources, such as a patient's appointment with a physician, lab tests, doctor's analysis and examination, etc. Due to this, data accessibility can be limited and the process becomes complicated for data collection, storage, and analysis. However, any data should not be ignored as all components of the data can have a significant impact on a patient's diagnosis and progressions. Therefore, the data need to be collected. Another issue is the incomplete or unstandardized data, inaccurate or missing data in the medical records. Various formats, for example, can be used to capture pieces of data in various sources. Without normal clinical terminology, data mining in

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 08, August-2020

the healthcare sector is also extremely difficult. A further barrier to effective data mining is poor mathematical characterization and non-canonical nature of such high volume, complex and heterogeneous data. There are also other important medical data issues, such as data control, ethical problems, social and legal concerns, etc. Another issue is that the data mining results can reveal various important and interesting trends which may be useless due to large data. Another requirement for successful data mining application is knowledge in the domain area, together with a proper understanding of data mining techniques. In addition, significant investment is needed in terms of time, resources and effort to improve data mining technology. The data entry should be systematic and appropriately stored for future use. The main requirement is thorough planning, technological preparation work, awareness of the technology's effectiveness and its use, collaborative and cooperative work by everyone involved in data mining.

VI. CONCLUSION:

The paper aimed on reviewing the works carried out on data mining in medical field. It is observed that data mining took on an evolving position because of the need to use data mining techniques in healthcare. Exploring knowledge from medical data is such a risky task as the data found are noisy, massive and also irrelevant. Over the last decades, health data holders have paid greater attention to data mining techniques, as these techniques can help them obtain very valuable knowledge. Such knowledge can be used to enhance various health services. It is also observed that a combination of more than one data mining tool come in handy in exploring the knowledge of medical data. Developing efficient data mining tools and techniques for an application could reduce cost and time constraint in terms of human resource and expertise.

REFERENCES:

[1] Abdulsalam Yassine, Shailendra Singh, Atif Alamri. "Mining Human Activity Patterns From Smart Ho me Big Data for Health Care Applications", IEEE Access, 2017.
[2] Pooja H, Dr. Prabhudev Jagadeesh M P , A Collective Study of Data Mining Techniques for the Big Health Data available from the Electronic Health Records, 2019 IEEE.
[3] Bindiya M K, Dr. Ravikumar, Dr. Mohan H S, "Monitoring the appliances used in health care using Medical Big Data", Proceedings of the International Conference on Inventive Computation Technologies(ICICT-2018).
[4] Damir Imamovic, Elmir Babovic, Nina Bijedic, "Prediction of mortality in patients with cardiovascular disease using data mining methods",19th International Symposium Infotech-Jahorina,18-20 March 2020.
[5] M. Durairaj, V. Ranjani, "Data Mining Applications In Healthcare Sector: A Study", International Journal of Scientific & Technology Research" Volume 2, October 2013.
[6] Geetha Guttikonda, MadhaviLatha Pandala, Madhavi Katamaneni, "Diabetes Data Prediction Using Spark and Analysis in Hue Over Big Data", Proceedings of the Third International Conference on Computing Methodologies and Communication (ICCMC 2019).
[7] Mohammad Hossein Tekieh1, Bijan Raahemi, "Importance of Data Mining in Healthcare: A Survey", IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2015.
[8] Sandeep Yadav, Aman Jain, Deepti Singh, "Early Prediction of Employee Attrition using Data Mining Techniques", IEEE, 2018.
[9] D.Usha Rani, " A survey on Data Mining Tools and Techniques in Medical Field", International Journal of Advanced Networking & Applications (IJANA), Volume: 08, Issue: 05 Pages: 51-54 (2017) Special Issue, TECHSA-17.
[10] Alramzana Nujum Navaz, Elfadil Mohammed, Mohamed Adel Serhani and Nazar Zaki. " The Use of Data Mining Techniques to Predict Mortality and Length of Stay in an ICU", 12th International Conference on Innovations in Information Technology (IIT), 2016.
[11] Ogundele I.O, Popoola O.L, Oyesola O.O, Orija K.T, "A Review on Data Mining in Healthcare", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 7, Issue 9, September 2018.
[12] Sayali Sunil, Abhishek Jamadar, Siddharth Dudugu Tandel, "A Survey on Text Mining Techniques", 5th International Conference on Advanced Computing & Communication Systems (ICACCS 2019).
[13] P Amrutha Valli, KRS.Prav allika, M Uma, Sasikala T, "Tracing out Various Diseases by Analyzing Twitter Data Applying Data Mining Techniques", International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS-2017).
[14] Pooja.H, Dr. Prabhudev Jagadeesh M P, "A Collective Study of Data Mining Techniques for the Big Health Data available from the Electronic Health Records", IEEE, 2019.
[15] Gaspard Harerimana, Beakcheol Jang, Jong Wook Kim, Hung Kook Park ,"Health Big Data Analytics: A Technology Survey", Department of Computer Science, Sangmyung University, oul, South Korea, IEEE. Translations, VOLUME XX, 2018.
[16] Pooja H , Dr. Prabhudev Jagadeesh M P , "A Collective Study of Data Mining Techniques for the Big Health Data available from the Electronic Health Records", IEEE Xplore, 2019 .
[17] Wencheng Sun, Zhiping Cai1 , Fang Liu , Shengqun Fang , Guoyan Wang, "A Survey of Data Mining Technology on Electronic Medical Records", 19th International Conference on e-Health Networking, Applications and Services, 2017 IEEE.
[18] Ritu Chauhan, Rajesh Jangade, "A Robust Model for Big Health care Data Analytics", 6th International Conference - Cloud System and Big Data Engineering, 2016.
[19] Mario W. L. Moreira , Joel J. P. C. Rodrigues , Senior Member, Valery Korotaev, Jalal Al-Muhtadi, and Neeraj Kumar, "A Comprehensive Review on Smart Decision Support Systems for Health Care", 2019 IEEE.