ISSN:0975-9646

Medical Data Mining Life Cycle and its Role in Medical Domain

Surabhi Thorat[1], Seema Kute[2]

[1&2]*MCA Department,*
Marathwada Institute of Technology(E)
Aurangabad(M.S), India

***Abstract*— Data mining is defined as a computational process of pattern detection from large datasets that leads to the extraction of information that is valuable and often 'hidden' .Data mining plays a vital role in decision making and for predicting the future trends of market. In recent years data mining is becoming a buzzword in the field of medical science. In this paper, I highlighted the Medical Data Mining Life Cycle which represents the complete phases that are involved in making the best possible solutions for the issues arises in healthcare sectors, it also presents a brief introduction of data mining approach in the field of medical science and focuses on the challenges in healthcare sector and the tools that are used in data mining.**

***Keywords*— Data Mining, Medical Data Mining Life Cycle, Data Mining Challenges, Data Mining Tools.**

I. INTRODUCTION

The term data mining first appeared in the 1960s while before that, statisticians used the terms "Data Fishing" or "Data Dredging" to refer to analysing data without an a-priori hypothesis. The most important objective of any data mining process is to find useful information that is easily understood in large data sets [1].

Data mining is the process of finding useful information from huge data sets. It is a technique that focuses on getting the patterns with minimal user input and efforts. It is "The science of extracting useful information from large database". It is one of the tasks in the process of knowledge discovery from the database. In other words, it is a combination of statistics and artificial intelligence which allows us to explore for patterns in large datasets [2].There are two primary goals of data mining tend to be prediction and description. Prediction involves some variables or fields in the dataset to predict unknown or future values of other variables of interest. On the other hand description focuses on finding patterns describing the data that can be interpreted by humans [3].

According to the WHO Global Health Observatory Analysis Ischaemic heart disease, stroke, chronic obstructive lung disease and lower respiratory infections have remained the top killers during the past decade. Chronic diseases cause increasing numbers of deaths worldwide. Diabetes caused 1.5 million (2.7%) deaths in 2012, up from 1.0 million (2.0%) deaths in 2000. Lung cancers (along with trachea and bronchus cancers) went up to become the 5th leading cause of death in 2012, killing 1.1 million men and 0.5 million women in 2012. Injuries continue to kill 5 million people each year. Road traffic injuries claimed about 3400 lives each day in 2012, about three-quarters of those were men and boys[4].

Data mining holds great potential for the healthcare industry that can reduce the mortality ratio. It enables health systems to systematically use the data and analytics to identify inefficiencies and best practices that improve care and reduce costs. Some experts believe the opportunities to improve care and reduce costs concurrently could apply to as much as 30% of overall healthcare spending. This could be a win/win overall, but due to the complexity of healthcare and a slower rate of technology adoption, our industry lags behind these others in implementing effective data mining and analytic strategies [5].

In recent years, data mining is tremendously used in medical domain like, DNA, genetics, medicine, biomedical. Data mining tool helps in large extent to increase the performance in the sense of diagnostics, prevention and treatment of the diseases. In medical domain data mining works on the bases of data that has been already collected and find the best possible solution by analysing and identifying the frequent pattern or trends of past data. In medical science past experience plays a vital role in diagnosing any new situation.

Basically the aim of using Data mining technique in medical domain is to facilitate hospitals, clinics, physicians, and patients by adopting new technologies, which will help in early detection of life threatening diseases, reducing treatment costs and increasing the survivability of the patient. The patient's data might contain potential facts or details about the possibility of diseases that can occur in a patient. This knowledge can be a great help for better diagnosis and treatment for future medical cases. This can only be possible if certain services can be provided in effective way:

- Facilitate with safe healthcare treatments

- Implement scientific medical knowledge to provide healthcare services.
- Providing various healthcare treatments based on the patient's needs, symptoms and preferences
- Time reduction for the medical treatment
- Health determinants
- Health outcomes (e.g., mortality, morbidity, disability, well-being, and health status)

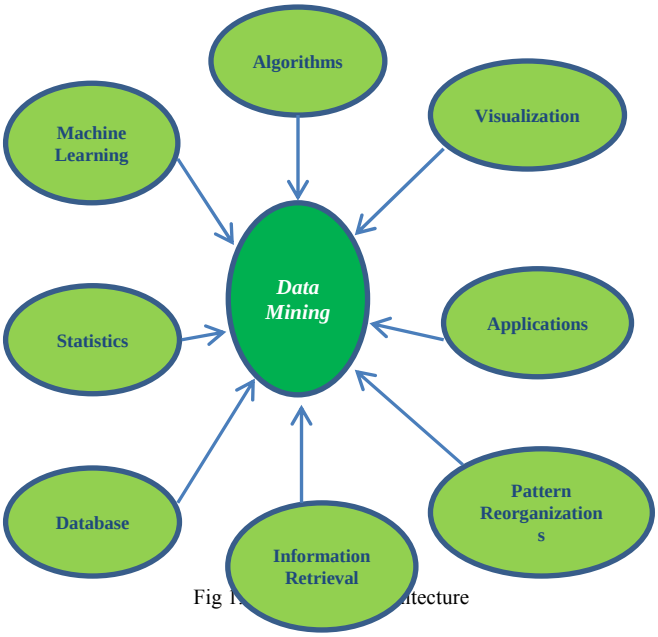

Fig 1 [illegible]tecture

II. Issues that arises the need of Data Mining

- To handle huge and complex data generated by healthcare organizations
- Traditional techniques infeasible for raw data.
- Information generated through data mining can be useful to all stakeholders in health care, including patients by identifying effective treatments and best practices
- There is often information "hidden" in the data that is not readily evident
- Human analysts may take weeks to discover useful information
- To reduce insurance fraud and abuse of insurers.
- Research advancements
- Improved treatments and medical devices

III. Medical Data Mining Life Cycle

Data mining provides great benefits to health care industry. The Medical Data Mining Life Cycle shows each and every step of finding out the best possible treatment or solution for patients that can save their lives. Medical Data Mining Life Cycle goes through major six regress phases to achieve its goal.

A. Data Collection

The very first phase of Medical Data Mining Life Cycle starts with the collection of huge amounts of data generated by healthcare transactions that is too complex and voluminous to be processed and analysed. Data from heterogeneous locations were collected like hospital, laboratories, radiology and administration.

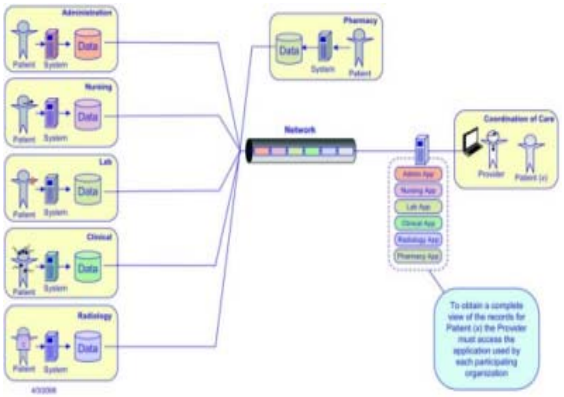

Fig 2. Electronic Medical Record Collection[6]

B. Association

In this phase an association between two or more symptoms that are often of the same type is formulates to find specific pattern. Associativity in information is helpful to uncover relationships between seemingly unrelated data in present in electronic medical record or other information repository. Basically association rule establishment lies between two concepts, an "if" and a "then". An "if" is an item found in the data. A "then" is an item that is found in combination with the "if". Association rules are useful for analysing and predicting similar symptoms for any problem related to health. There is a number of cases where associativity arises .Like it is observed that a person working in chemical industry is more prone to lung disorders or a person working in coal industry may suffer from respiratory issues.

C. Classification

Classification consists of predicting a certain outcome based on a given input. In order to predict the outcome, the algorithm processes a training set containing a set of attributes and the respective outcome, usually called goal or prediction attribute. The algorithm tries to discover relationships between the attributes that would make it possible to predict the outcome. Next the algorithm is given a data set not seen before, called prediction set, which contains the same set of attributes, except for the prediction attribute – not yet known. The algorithm analyses the input and produces a prediction. The prediction accuracy defines how "good" the algorithm is. [7]

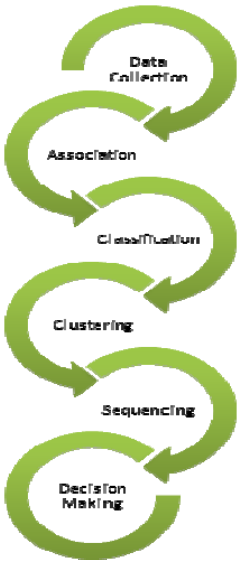

Fig 3. Medical Data Mining Life Cycle

This phase classify the collected information according to our motive like etiological factors, investigation purpose, drug treatment plans and results. Example, the etiological information collected from lung cancer patients can be classified on the basis of duration of toxic inhalation, smoking habit, type of exposure, number of exposure, age of patient etc.

Base Classifiers are:

- Decision Tree based Methods
- Rule-based Methods
- Nearest-neighbor
- Neural Networks
- Naïve Bayes and Bayesian Belief Networks
- Support Vector Machines

D. Clustering

The collected data is transformed in the form of cluster by using computer graphics; it is easy to locate trends of any particular disease. The clustering problem has been identified in numerous contexts and addressed being proven beneficial in many medical applications. Clustering the medical data into small with meaningful data can aid in the discovery of patterns by supporting the extraction of numerous appropriate features from each of the clusters thereby introducing structure into the data and aiding the application of conventional data mining techniques [8].These identified trends may helpful in preparation of prediction system. Information collected from any disease onset and duration of exposure, clustering pattern guide us in future regarding prevention strategies.

E. Sequencing

The data pattern sequencing is next step in module preparation. Mostly the interaction between the hospital and the patients is temporary in nature, in this phase symptoms sequentiallity is developed. Diagnosis and correlated procedure is procedure is prepared. It helps to mine sequential patterns or model temporal characteristics are likely to be central in many data mining efforts. The pattern sequencing can be prepared with the help of various software packages available in market or freeware.

F. Decision Making

The last step in Medical Data Mining Life Cycle is decision making using decision tree that provides the solution to many medical treatments. There are many benefits in data mining that decision trees offer:

- Self-explanatory and easy to follow when compacted
- Able to handle a variety of input data: nominal, numeric and textual
- Able to process datasets that may have errors or missing values
- High predictive performance for a relatively small computational effort
- Available in many data mining packages over a variety of platforms
- Useful for various tasks, such as classification, regression, clustering and feature selection[9].

IV. CHALLENGES IN HEALTH SECTOR

In medical research, data mining starts with a hypothesis and then the results are adjusted to fit the hypothesis. This diverges from standard data mining practice, which simply starts with the dataset without an apparent hypothesis[10].The major challenges of data mining in medical domain based on these issues:

A. Data Quality

The biggest challenge of data mining in health sector is its data quality. It is difficult to acquire the precise and complete healthcare data. Health data is complex and heterogeneous in nature because it is collected from various sources such as from the medical reports of laboratory, from the discussion with patient or from the review of physician. For healthcare provider, it is essential to maintain the quality of data because this data is useful to provide cost effective healthcare treatments to the patients. Health Care Financing Administration maintains the minimum data set (MDS) which is recorded by all hospitals. In MDS there are 300 questions which are answered by the patients at check-in time. But this process is complex and patients face problem to respond the entire questions. Due to this MDS face some difficulties such as missing information and incorrect entries. Without quality data there is no useful results. For successful data mining, complication in medical data is one the significant hurdle for analyzing medical data. So, it is essential to maintain the quality and accuracy data for data mining to making effective decision.

The correct information can only be achieved, if the quality data is available for mining.

The data quality can be measured in terms of:

- Comprehensiveness-To identify that complete information of patient is available or not?

- Perfection-It determines the accuracy of electronic data present about patient.
- Similarity-To find similarity between elements of electronic records.
- Credibility-Does the knowledge can be discovered from the electronic health record**[9].**

B. Privacy

One of the challenges in these circumstances is privacy. Most of the patients do not want to disclose their health data. So, the Health Maintenance Organization and Health insurance Organization are not distributing their data for preserving the privacy of patient. This poses hurdle in the fraud detection studies in health insurance. The hospitals that wish to maintain the data ware houses are concern about maintain the privacy of patient. This can lead to the problem, that potentially meaningful information can remain undiscovered. To overcome this situation certain privacy-preserving policies must be prepared.

C. Expensive Implementation

Implementing and maintaining a data warehouse in organization at the start level is expensive in terms of budget. Because before applying data mining techniques in healthcare data it is essential to collect and record the data from different sources into a central data warehouse which is a costly and time consuming process. Faulty data warehouse design does not contribute to effective data mining. Small sized hospital cannot effort such huge expense of implementing data warehouse.

D. . Credibility

Data mining of medical data requires specific medical knowledge as well as knowledge of Data mining technology both of them complement each other .Less ratio of any one of will not provide the best outcome.

E. Lack of Integration

Electronic health record is distributed among hospitals, insurance companies and government departments. Due to this data integration and data mining put on stack, in terms of the reliability that can be achieved knowledge discovery of results and the semantics of a derived rule. It arise the need of uniform standards to integrate data from heterogeneous systems.

V. DATA MINING TOOLS TO PREDICT DISEASE

Various data mining tools are available that can be used to predict the accuracy level in different health problems. The data mining tools are broadly classified into 3 categories.

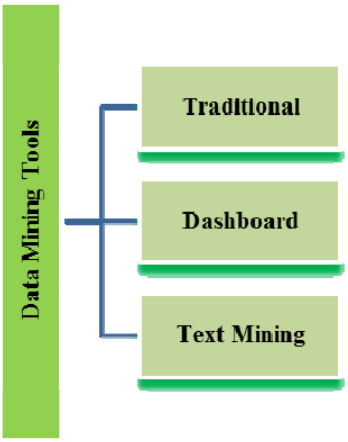

Fig 4. Data Mining Tools Classification

A. Traditional Data Mining Tools-

Traditional data mining tools help companies establish data patterns and trends by using a number of complex algorithms and techniques. Some of these tools are installed on the desktop to monitor the data and highlight trends and others capture information residing outside a database. The majority are available in both Windows and UNIX versions, although some specialize in one operating system only. In addition, while some may concentrate on one database type, most will be able to handle any data using online analytical processing or a similar technology.

B. Dashboards-

Installed in computers to monitor information in a database, dashboards reflect data changes and updates onscreen often in the form of a chart or table that enables the user to see how the business is performing. Historical data also can be referenced, enabling the user to see where things have changed. This functionality makes dashboards easy to use and particularly appealing to managers who wish to have an overview of the company's performance.

C. Text-mining Tools.-

The third type of data mining tool sometimes is called a text-mining tool because of its ability to mine data from different kinds of text from Microsoft Word and Acrobat PDF documents to simple text files, for example. These tools scan content and convert the selected data into a format that is compatible with the tool's database, thus providing users with an easy and convenient way of accessing data without the need to open different applications. Scanned content can be unstructured (i.e., information is scattered almost randomly across the document, including e-mails, Internet pages, audio and video data) or structured (i.e., the data's form and purpose is known, such as content found in a database). Capturing these inputs can provide organizations with a wealth of information that can be mined to discover trends, concepts, and attitudes.

VI. CONCLUSION

Data mining play a vital role in medical domain by which better diagnosis and treatments can be provided to the patients. In past years it has been observed that mortality radio is rising tremendously, this can be reduced by making use of current techniques available in the field of technology. Data mining is evolved with a great scope in better diagnosis and outcomes in the medical science. In this paper, I focused about regress process of decision making and diagnoses by applying the data mining concepts and challenges that is to be considered for better results. Different data mining tools can be utilized in medial domain. On the basics of it I highlight Medical Data Mining Life Cycle .It shows the phases that should be followed to get the effective and timely results so that the death ratio would be reduced.

REFERENCES

[1] http://en.wikipedia.org/wiki/Data_mining
[2] Eirini Papageorgiou,loanna Kotsioni,Athena inos "Data Mining:A New Technique In Medical Research"
[3] S.Vijiyarsani,S.Sudha,"Disease Prediction in Data Mining Technique – A Survey", International Journal of Computer Applications & Information Technology Vol. II, Issue I, January 2013
[4] http://www.who.int/gho/mortality_burden_disease/en/
[5] "http://www.healthcatalyst.com/data-mining- in-healthcare" , by David Crockett, Ryan Johnson and Brian Eliason on May 28, 2014. Posted in Analytics in Health , Predictive Analytics
[6] Electronic Health Records Overview(April 2006), National Institutes of Health National Center for Research Resources
[7] http://courses.cs.washington.edu/courses/csep521/07wi/prj/leonardo_fabricio.pdf
[8] F. H. Saad, B. de la Iglesia, and G. D. Bell, ― A Comparison of Two Document Clustering Approaches for Clustering Medical Documents, Proceedings of the 2006 International Conference on Data Mining (DMIN-06), 2006.
[9] Lior Rokach , Oded Maimon , Data Mining with Decision Trees Theory and Applications.
[10] DATA MINING IN HEALTHCARE: CURRENT APPLICATIONS AND ISSUES. By Ruben D. Canlas Jr