Data mining in healthcare data management-A survey

Vartika Punjabi[1], Shashwat Mishra[2], Manikandan K[3]
vartikapunjabi1@gmail.com[1], shashwat.mishra2015@vit.ac.in[2] , kmanikandan@vit.ac.in[3]
School of computer science and engineering, Vellore Institute of Technology

ABSTRACT
In this study, we focus on collecting relevant information to clearly understand the relationship between data mining and healthcare management. The healthcare datasets now include health records, insurance claims, health surveys and many other sources. With the exponential growth of healthcare data, being collected every day, data mining tools have become essential for analyzing, maintaining and making future predictions. They prove to be of great importance for identifying special characteristics of health data i.e. identifying new health conditions and various other applications. However these techniques still face a very complex challenge of interpretation, data quality and piracy. This study also focuses on how public health data can be used along with various data mining techniques for effective decision making. The aim is to promote data mining analysis and continuous evaluation of models on certain fixed parameters to make the optimum use of public health data.

Keywords *- data mining, healthcare data management, public health data, decision making*

1. INTRODUCTION

Over the past decade we have observed a drastic shift from paper based record systems to electronic systems. Electronic systems are more efficient and thus offer benefits for everyone involved. [2]However with the increase in this data has raised many concerns about the feasibility in maintaining these data sets but has also created a new horizon of opportunities, to identify the information in these datasets. Data mining is thus gaining popularity in organizations, who seek to gain information that can directly influence the efficiency of the organization in terms of quality, expenses, services etc. Due to this increasing demand data mining algorithms are evolving and being developed to more reliable information.

[5]Healthcare data is a huge dataset ranging from online surveys, insurance claims, electronic medical records(EMR) etc. thus the majority data is being collected by hospitals, clinics and various other health provides. However this data is not easy to analyze, being unstructured, and thus requires various data mining techniques for cleaning, organizing , finding hidden trends to finally provide some physical benefits to the organization using predictive analytics. [4]This complex data set can be classified broadly as structured and unstructured. Structured data includes EMR systems claims system, revenue cycle, public data, benchmark data whereas unstructured data includes prescriptions, images, videos, social media. In this age of fast digitization, Indian population are adopting the global trends, like mobile health, remote patient monitoring etc., at a much faster pace. These new datasets have become more distinctive from the traditional datasets in terms of volume, variety and velocity thus becoming "big data".

[3]The pubic healthcare in India include a set of facilities in which some are owned by the state government and some by the central government. The union of health and family welfare focuses on efficient and effective administration by the state and local health. [1]It has been observed that past techniques for data analysis have been statistical and data mining techniques are secondary tools. In this study, we demonstrate through certain evidences and explanations how data mining can play a key role in public healthcare data interpretation. It also focuses on identifying various techniques and models used for classification along with the process involved for efficient data mining.

2. PROBLEMS IN HEALTHCARE

From the related work it is evident that most of the developments in the healthcare field are concentrated on ongoing care management whereas very few studies are focused on insurance claim generation and management and hospitalization utilization. Ongoing care management includes disease prediction and survival analysis. Disease prediction and identification include various studies for detecting cancer, hypertension, diabetes, tumors, liver disease, thyroid, SARS etc. This data is further used for conducting a survival analysis after some disease or incident. So far survival analysis has been successfully implemented for kidney failure, cancer patients and burn victims. Other research studies are exploratory in nature like association rule analysis has been done to identify patterns between certain medical conditions in a particular geographical area or to establish relationship between environmental chemicals and related medical conditions.

Hospital utilization management primarily focus on reducing the length of stay of patients along with a low re-admission rate. Few studies have been focused on this issue by analyzing readmission of heart failure patients, predicting the chances of readmission based on a fixed set of parameters for body vitals at the time of discharge. Whereas several studies have been solely focused, using data mining techniques, on predicting the length of stay of certain patients. Claim management is gaining popularity for fraud detection and need detection of the consumers. Data mining techniques are also gaining popularity in other fields like cost prediction for cancer patients and disease outbreak.

3. DATA MINING TECHNIQUES

Data mining is a strong tool for extracting useful information from complex datasets and to evaluate relationships between the attributes of the data. It origins from intertwining concepts of statistics and machine learning which now has been listed as one of the top 10 leading technologies to change the world, after witnessing an exponential growth in the past 2 decades.

Data mining techniques can be broadly classified in two categories namely supervised and unsupervised learning. Supervised learning generates predictive rules as a model to classify the target records whereas unsupervised learning attempts to measure similarity of records and discover patterns. Clustering and association are key techniques used in unsupervised learning on the other hand classification is the key activity in supervised learning.

3.1 Classification

This technique is used to classify data into target attributes using the input attributes. Some of the techniques used in this process are decision trees, neural networks, k-nearest neighbours, support vector machine, Bayesian methods. Several researches show that decision tree is the most widely used classification technique for healthcare data. It is observed to be used for analyzing microarray data, diagnosing skin diseases, performance of different classifiers on cancer datasets, predicting cost of healthcare services, identifying healthcare coverage and predicting patient status.

3.2 Clustering

It is a unsupervised technique used when very less information is available about the data objects involved in the population. It simulates clusters of data objects that reflect similarities to each other. Hence no predefined classes are used. Some techniques included partitioned clustering, hierarchical clustering and density based clustering.

3.3 Association

This technique highlights the relationship of attributes in a dataset. It is widely used to detect the relationship between diseases. A classifier is built using the identified rules and the main attributes.

Data mining poses some very important gains over traditional methods by using statistics along with machine learning, artificial intelligence and visualization. Due to flexibility data mining includes categorical attributes along with various other heuristics for analysing real world issues. It also has the ability to handle numeric data as well as the categorical attributes – e.g. Race, gender, diagnosis code etc. Data mining is highly efficient if we aim to discover new diseases since it does not consider any particular hypothesis instead tries to find information by exploring the given dataset.

Decision making is a process of identifying an selecting among various alternatives by adding a certain priority to the choices. It aims to reduce the uncertainty among the available alternatives and make a well informed decision. Decision making involves three basic levels – strategic tactical and operational. Strategic decision are taken by keeping in mind the organization's mission and vision for long term benefits. Tactical decisions are made during the implementation of strategic decisions. They are meant to manage performance for achieving the devised strategies. Operational decisions include all day-to-day operations of the organisation. Health decision tools are used to make smart choices in terms of medical tests, treatments, surgeries, reducing costs etc. Integrating data mining with decision making can enhance and improve the decision support system, functions and performances, especially where a huge dataset is involved.

4. APPLICATION OF DATA MINING IN HEALTHCARE

4.1. Clinical decision making

Patients visiting healthcare facilities will be analyzed by clinicians to analyze their issue or infection. In spite of the fact that the medical professionals do all their best in distinguishing the reasons for each side effect in the patient, the nature of this examination is trial and now and then the judgments may turn out badly. Information mining procedures can help the specialists in the field to get a second assessment for most judgments, particularly to ensure the infection isn't under-evaluated amid determination

4.2. Biomedicine and genetics

Some particular diseases are contemplated in the biomedical and sub-atomic dimension, along with the clinical dimension. As the measure of extricated biomedical and sub-atomic information are expanding, it can enable analysts to research the impacts of hereditary qualities on various infections in smaller scale level (sub-atomic examination); in this way, we have isolated this area from the populace well being segment which takes a gander at the pattern of illnesses in large scale level. Having said that, in microarray information investigation, grouping methods have gotten more consideration contrasting with order and relationship as there are not a great deal of data accessible about qualities, as opposed to wellbeing conditions and sickness side effects that a ton of data are known.

4.3. Population health

Disease transmission specialists and other wellbeing investigators concentrated on the predominance of ailments are keen on distinguishing the examples, patterns, and reasons for spreading a particular infection over a populace. For these investigations, they consider distinctive hazard components and wellbeing determinant, including early-life, way of life, and socio-statistic.

4.4. Health administration and policies

One of the enormous difficulties in the zone of wellbeing organization is taking care of protection plans and their system. Generally every nation or state has its own particular component. Be that as it may, one issue that all medical coverage organizations (regardless of it is open or private) manage is protection extortion. Information mining has been connected to identify protection extortion in which the patients, specialists, or medical clinics guarantee medicates that were a bit much or techniques that halfway or completely did not really occur.

4.5. Health big data

"Big Data" has turned into a realized term alluding to a gigantic accumulation of information to the point that the customary information investigation methods don't work productively. Notwithstanding the volume of information, speed (spilling information) and assortment (semi-organized and un-organized) are significant worries in "huge information" as well. Because of actualizing EHR and comparative frameworks to gather electronic wellbeing and therapeutic information, notwithstanding the information being gathered by means of online protection cases and wellbeing reviews, a lot of information has been gathered by wellbeing associations which is expanding regularly. Information mining methods can consider the entire populace into the investigation which can give data about points of interest, subtleties, and furthermore minorities.

4.6. Health condition mysteries

Not a great deal of data is accessible about qualities' effects on wellbeing issue. In this manner, exploratory techniques can essentially assist the scientists with identifying the connections between the characteristics and furthermore bunch the gatherings of contributing qualities. Expressive information mining has begun developing here with the endeavor of going from explicit (information) to general (learning) about the impacts of qualities on wellbeing conditions. In the following stages, characterization and forecast models can likewise be worked for the found gatherings to almost certainly foresee the wellbeing issue utilizing quality articulation designs saw in patient's connected organs.

5. CHALLENGES OF DATA MINING IN HEALTHCARE

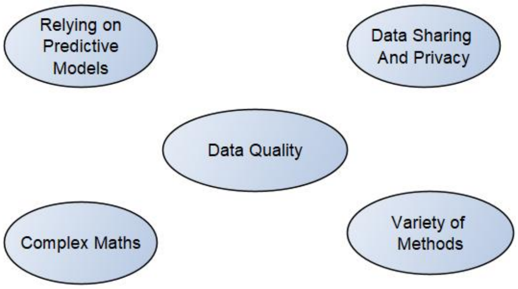

Fig.1. Various challenges faced by data mining in healthcare

5.1. Data quality

To accomplish reliable and useful data from a data mining process, it requires information with great quality. Wellbeing information is normally gathered from various sources with very surprising set-ups and database structures which makes the information mind boggling, filthy, with a great deal of missing information, and distinctive coding gauges for similar fields. For example, albeit hazardous penmanship styles are not any more relevant in EHR frameworks, the information gathered through these frameworks are not principally assembled for systematic purposes and contain numerous issues – missing information, error, miscoding – because of clinicians' outstanding tasks at hand, not easy to use UIs, and no legitimacy checks by people.

5.2. Data sharing and privacy

Since the health data contains personal health information (PHI), there will be legitimate troubles in getting to the information because of the danger of attacking the protection. This issue puts a major hole between the gathered information and the information expert, and associating these two is once in a while not extremely simple. Wellbeing suppliers are not normally alright with imparting their information to the examiners to maintain a strategic distance from any hazard that dangers the security of patients. Then again, setting up a protected framework to accumulate information from various sources is very tedious and costly [1]. No entrance to information basically implies no contribution to the information mining procedures, and in this manner, no examination and result data.

5.3. Relying on predictive models

There ought not be unlikely desires from the developed information mining models. Each model has an exactness. At the point when a prescient model for diagnosing a medical problem – for example diagnosing the kind of thyroid organ – is constructed, it more often than not does not have a precision of 100%. Contingent upon the measure of model's precision and the significance of the choice being made utilizing that demonstrate, we ought to choose the amount we can depend on the results of this information mining model. Particularly when we are managing clinical prescription choices dependent on the results of one or multi information mining thinks about, think about that it is unsafe to possibly depend on the prescient models when settling on basic choices that straightforwardly influences the patient's life, and this ought not be normal from the prescient model.

5.4. Variety of methods and complex maths

As the fundamental math of practically all information mining procedures is mind boggling and not in all respects effectively justifiable for non-specialized colleagues, consequently, clinicians and disease transmission specialists have normally liked to keep working with conventional measurements strategies. They comprehend the p esteem much better contrasting with information mining's estimation techniques, for example, accuracy, affectability, explicitness, and ROC bend. Moreover, as there are a wide range of information mining strategies and techniques, it is troublesome for clinicians to get acquainted with all the distinctive techniques and effectively select the correct one.

6. PROPOSED CONCEPTUAL MODEL

The sole purpose of the proposed conceptual model is to facilitate the integration of data mining and decision making process in public healthcare system. The reference model was modified according to the requirements of data mining model based on public healthcare system. The methodology of the model is discussed as follows

The implementation of the model begins with creating a basic understanding about the various domains of public healthcare, and setting objectives to improve wider determinants of health, health improvement, health protection and preventing premature mortality. The information is then collected at national, state and district levels either from the institutions offering public healthcare or from the general population. The data set should include large scale demographic surveys, performance statistics, vital indicators, national reports and agencies involved in such activities. For data analysis relevant and current data should be considered along with design, size, range, completeness, confidentiality, availability, etc. The collected data, stored on a local database, undergoes certain enhancements which involves more user friendly names of the variables, decoding any unique set of values etc.

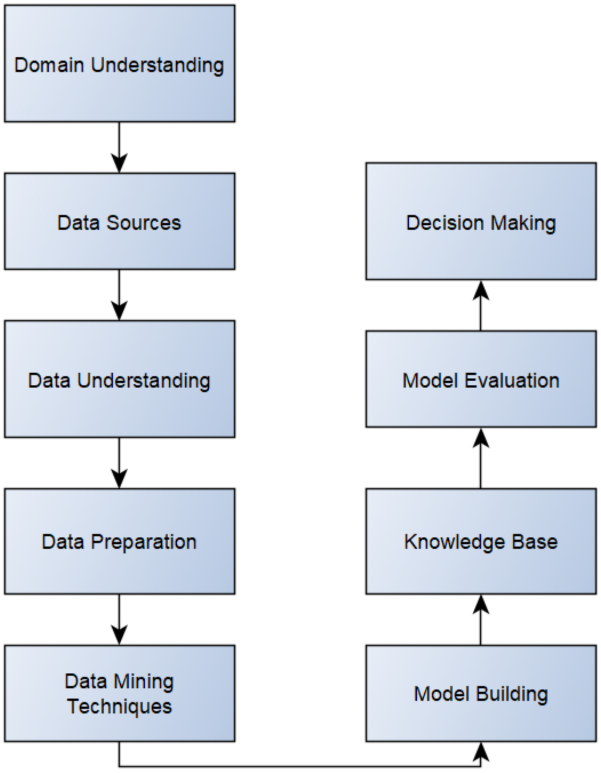

Fig.2. Proposed conceptual model

After identifying the data set format and the target issue to be addressed various data mining techniques like association rules, clustering or segmentation, classification are used. Time series data mining can also be used for predicting healthcare trends in a given region. From many existing algorithms, different algorithms can also be applied to the same healthcare dataset to carry out complex tasks. The model is selected on the bases of level i.e. national, state or district or types of decisions, strategic, tactical or operational model. Knowledge is obtained through a set of guidelines issued and extracted from the dataset. Depending on the task data extraction could be either predictive or descriptive. Quality of the knowledge will determine the usefulness and validity of the proposed model. The selected model is then evaluated several times to find the ideal and most efficient model to represent the data. Many iterations are performed to check for changes in the current dataset to evaluate performance on the basis of the training dataset. Since healthcare data is heavily unstructured the data modeling task becomes even more complex. It also shapes the future mission of the organization, the decisions include mapping public healthcare facilities, capacity building, hiring extra staff, scheduling available resources, routine decisions, emergency level situation etc.

6. SCENARIO IN INDIA

It is evident that through all the studies that geographical region has a key role to play in analysing any healthcare dataset. In terms of digitizing medical records India lags behind from many other countries. In spite of relaxed privacy norms in India only one data set is publicly available in UCI. The dataset is fairly simple and structured. From the current literature study it is found that the current researches are mostly focused on ongoing care management especially heart related ailments whereas relatively less or almost no work has been done in fraudulent claim detection or disease outbreak detection. Several machine learning techniques can automate diagnoses with a reasonable accuracy. Any development in these techniques will be extremely beneficial for the increasing rural population and staggering doctor to population ratio.

7. CONCLUSION

By analyzing the dataset through its several attributes and then extracting the hidden information can lead to the development of many useful and practical solutions. Since the data through electronic records systems is growing exponentially it has become a need to come with optimum solutions for data management and extracting information to improve health services and deliveries, identifying relationship between diseases and making the organization very cost efficient. Data mining plays a key role in this study as traditional statistical models cannot handle this huge amount of data at once. Data mining is also capable of identifying patterns that can cause over budgeting, classifying microarray data for unknown health issues, predicting insurance frauds and high risk patients. However the implementation of these techniques faces many challenges such as algorithm performance, information reliability, data quality and variety of complex methods.

The aim of the conceptual model was to offer a easy to understand approach for healthcare officials to exploit the available data mining tools by integrating them with their decision making process. It also provides a platform for applying data mining techniques at different management levels and can be implemented in any service sector.

Also this paper presents the current relationship of data mining and healthcare in India. India is lagging heavily in making the data available for research collaborations. This can serve as a basis for future researchers to create close collaborations between the computer scientists and medical officials along with creating ample amount of research opportunities with the growing amount of unstructured data.

REFERENCES

[1] Johannes K. Chiang, "Multidimensional Data Mining of Association Patterns in various Granularities for Healthcare Service Portfolio Management", IEEE Computer Society, pp. 525-531, 2007
[2] Hui Yang and Erhun Kundakcioglu, "Healthcare Intelligence: Turning Data into Knowledge", Trends and Controversies, IEEE Computer Society, pp. 54-68, 2014

[3] Amit Kumar Das, Aman Kedia, Lisha Sinha, Saptarsi Goswami, Tamal Chakrabarti and Amlan Chakrabarti, “Data mining techniques in Indian Healthcare: A Short Review”, International Conference on Man and Machine Interfacing (MAMI), 2015

[4] Peng Zhang, Shang Hu, Jing He, Yanchun Zhang, Guangyan Huang and Jiekui Zhang, “Building Cloud-Based Healthcare Data Mining Services”, IEEE International Conference on Services Computing, pp. 459-466, 2016

[5] Anand Sharma and Vibhakar Mansotra, “Data Mining Based Decision Making: A Conceptual Model for Public Healthcare System”, International Conference on Computing for Sustainable Global Development (INDIACom), pp. 1226-1230, 2016