A survey on Data Mining Methodologies for Healthcare

Dr. Mahendra Tiwar¹, Ms. Papiya Mukherjee², Dr. Amarjeet Singh³

¹,²Assistant Professor, Department of Computer Science & Engineering, United College of Engineering and Research, Allahabad, Uttar Pradesh, INDIA
³Sr. Assistant Professor, Shriram Institute of Management and Technology, Kashipur, Uttarakhand, INDIA

ABSTRACT
Data Mining is one of the most motivating area of research that is become increasingly popular in health organization. This survey explores the various process models for data mining in health sector, in this paper we present a brief introduction of most popular methodology, CRISP-DM along with some extensions of this model for health care.

Keywords----- Data Mining, Process model, CRISP-DM

I. INTRODUCTION
In India healthcare is delivered through both the public sector and private sector. The public healthcare system consist of healthcare facilities run by central and state government which provides service free of cost or at a subsidized rates of low income group in rural and urban areas. With the indian economy enjoying a steady growth, (M.Radha,2011). Good health confers on a person or groups freedom from illness- and the ability to realize one’s potential. Health is therefore best understood as the indispensable basis for defining a person’s sense of well being. The health of populations is a distinct key issue in public policy discourse in every mature society often determining the deployment of huge society.

Emerging private sector is more focused on tertiary-level as well as preventive and diagnostic healthcare and is sensing a huge untapped opportunity in delivery of quality healthcare to the indian masses Data Mining in healthcare.

Data mining can be considered as a relatively recent developed methodology and technology, coming into prominence. It aims to identify valid, novel, potentially useful, and understandable correlation and patterns in data by combining through copious data set of sniff out patterns that are too subtle or complex for human to detect. In healthcare, data mining is gradually increasing popularity, If not by any case, becoming increasingly essential. Data mining application can greatly benefits all parties involved in the healthcare industry. In healthcare, data mining is becoming increasingly popular, if not, increasingly essential. Several factors have motivated the use of data mining in healthcare. While arriving at a conclusive medical decision, data mining support assumes a high significance, Analysis of digitized images of skin lesions to diagnose melanoma, computer assisted texture analysis of ultrasound images aids monitoring of tumor response to chemotherapy, are some of its applications.

1. Diagnosis and Treatment
3. Prediction of impatient length of stay
4. Customer Relation Management
5. Unhealthy insurance practice
6. Data quality and completeness
7. Obstacles for data mining in healthcare

II. PROCESS MODEL FOR DATA MINING IN HEALTH SECTOR

A. CRISP-DM
1. CRISP-DM (Cross Industry Standard Process for Data Mining) provides a standard for carrying out data mining activities in six phases. Details of these phases are as follows:
2. Business understanding: This phase highlights the understanding of the business activities and determine the initial plan of achievement of goals defining the success criteria.
3. Data understanding: In this phase data for business activities are collected and analyzed.
4. Data Preparation: The data preparation phase covers all the activities required for quality of
data, it involve choosing of table, classification, normalization, noise removal and sampling.

5. Modeling: This phase is about to select a suitable modeling techniques, algorithms, for selecting optimal technique or algorithm it perform iterative way until it is achieved the chosen model quality criteria.

6. Evaluation: This phase evaluates high-quality model, it is mandatory to carefully evaluate the model before deployment and make sure that business objectives are achieved.

7. Deployment: In this phase deployment of constructed model is performed.

B. CRISP-TDM

CRISP-TDM(Cross Industry Standard Process-Temporal and Multidimensional aspect) is an extension of Crisp-DM for improving clinical care. This model supports the process mining in clinical environment.

The first phase of this model is Business understanding phase. The objective will be to find out whether new pathophysiological behaviors can be determined in retrospective analysis of physiological data streams prior to the diagnosis of a selected clinical event of interest.

The second phase is Data understanding, with use of STDM(Service based multidimensional temporal data mining), the data are collected and contained within the data management layer of the architecture. In addition, several temporal abstraction against that data will have already been run. The rules for which are contained in the temporal rules table, and the derived patient data is contained within the temporal abstraction tables. With these processes Data Preparation is also completed. After this step, the next task is to establish a study data subset for process mining. Within STDM, it is performed by defining a new study is defined, the clinical condition event of interest is defined and, associated with that, the rules for the selection of patients to be included in that retrospective study.

C. CRISP-MED-DM

Another extension of CRISP-DM in healthcare is CRISP-MED-DM. In CRISP-MED-DM, the first phase of this extended model is problem understanding where Determine overall objective is divided in to define clinical objectives, define healthcare objective, and define success criteria. another dimensions of problem understanding are: Assess situation, determine data mining goals, and plan activities.

In the second phase data understanding, a new general task “prepare for data collection was introduced. Issues of transport, semantic and functional interoperability have been considered in this activity. Prepare for data collection, collect initial data, describe data, explore data, and verify data quality are dimensions of this phase.

The third phase of this model is data preparation some new general task are added like prepare data, extract data, integrate data, clean data, select data, construct data, and format data.

Fourth phase of this model is modeling the task of this phase are: select modeling technique, generate test design, prepare model for interoperable use, build model, and assess model.

Evaluation is fifth phase where evaluate results, review process, and determine next steps are the activities involved.

The sixth phase of this process model is deployment, the general task which involved are, plan deployment, review project, plan monitoring and maintenance, produce final report.

This technique is most important technique which is used in Knowledge Discovery in Database(KDD). KDD has different types of steps like Data cleaning, Data integration, Data selection, Data transformation, Data mining, Pattern evaluation, Knowledge presentation etc. There are different types of techniques used in Data mining project. These include Decision tree, Bayesian networks, Naive bayes, Neural networks etc. Decision tree-It is the most frequently used techniques of data analysis. It is used to classify records to a proper class and is applicable in both regression and associations tasks. In medical field decision trees specify the sequence of attributes. Such a tree is built of nodes which specify conditional attributes – symptoms X=[x1,x2,......xk],branches which show the values of S i.e. the h-th range for i-th symptom and leaves which present decisions Y=[y1,y2,......yk] and their binary values Zdk={0,1} . A sample decision tree is presented in the fig1.

![Decision tree applicable in medicine](image)

Naive Bayse- It is a simple probabilistic classifier, which is based on an assumption about mutual independency of attributes. The probabilities which is
applied in the Naïve Bayes algorithm are calculated according to the Bayes Rule, the probability of hypothesis \( H \) can be calculated on the basis of the hypothesis \( H \) and evidence about the hypothesis \( E \) according to the following formula:

\[
P(H|E) = \frac{P(E|H) \cdot P(H)}{P(E)}
\]

Neural Networks-In medical diagnosis the input to the neural network are the patient’s symptoms the set \( X \), and \( Y \) is the output of the diagnosis. There are 3 layers in neural networks: input layer, hidden layer, output layer. Hidden layer is the outcomes of the input layer. The condition between neurons have weights which is assigned to them. Their values are calculated with the use of back propagation algorithm. In hidden layers there are some nonlinear features are added to the network. The out layer may have more than one output node which predict the different diseases. In a single neuron there are many input layers and one output layer. The input and output values are issued with the use of combination and activation function.

![Fig 2 Single neuron](image)

A. Advantages of Data mining
- Predict future trends, customer purchase habits
- Help with decision making
- Improve company revenue and lower costs
- Market basket analysis
- Fraud detection

B. Disadvantages
- Great cost at implementation stage
- Possible misuse of information
- Possible inaccuracy of data

III. DATA MININ IN HEALTHCARE

Data mining applications are currently being applied to two main branches in health care and medicine: Healthcare decision support system, and policy planning/decision making. [5][6] A. Healthcare decision support system HDSS is an interactive Decision support system(DSS) Computer Software, which is designed to assist physicians and other health professionals with decision making tasks, such as determining diagnosis of patient data. The main purpose of modern HDSS is to help clinicians at the point of care. It means, a clinician would interact with a HDSS to help determine diagnosis, analysis, etc. of patient data. It is a decision-support system program that offers employees in-depth, objective, personalized, and current information on all healthcare conditions. Employees receive the information, tools, and support they need from integrated web, phone, and print based materials. This helps employees make more informed healthcare decisions while working with their own physician.[7] There are two main types of HDSS.
- Knowledge-Based
- Non Knowledge-Based

An example of how a HDSS might be used by a medicinal comes from the subset of HDSS (Healthcare Decision Support System), DDSS (Diagnosis Decision Support Systems). A DDSS would take the patients data and propose a set of appropriate diagnoses. The doctor then takes the output of the DDSS and point out which are relevant and which are not. Another important classification of a HDSS is based on the timing of its use. Doctors use these systems at point of care to help them as they are dealing with a patient, with the timing of use as either pre-diagnoses, during diagnoses, or post diagnoses. Pre-diagnoses HDSS systems are used to help the physician prepare the diagnoses. HDSS used during diagnoses help review and filter the physician’s preliminary diagnostic choices to improve their final results. And post-diagnoses HDSS systems are used to mine data to derive connections between patients and their past medical history and to predict future events. Features of a Knowledge-Based HDSS Most HDSS consist of three parts, the knowledge base, inference engine and mechanism to communicate. The knowledge base contains the IF-THEN rules. The inference engine combines the rules from the knowledge base with the patient’s data. The communication mechanism will allow the system to show the results to the user as well as have input into the system. Features of a non-Knowledge-Based HDSS Two types of non-knowledge-based systems are neural networks and genetic algorithm. Neural networks use nodes and weighted connections between them to analyze the patterns found in the patient data to derive the associations between the symptoms and a diagnosis. Genetic Algorithms are based on simplified evolutionary processes using directed selection to achieve optimal HDSS results. The HDSS features associated with success include the following:
- it is integrated into the health care workflow rather than as a separate log-in or screen.
- it is electronic rather than paper-based templates
- it provides decision support at the time and location of care rather than prior to or after the patient encounter.
- it provides (active voice) recommendations for care, not just assessments.
IV. SCOPE OF DATA MINING TECHNIQUES IN HEALTHCARE

Increasing computer based data analysis awareness, online educational availability and developing integrated learning approach among medical professionals will definitely helpful for accurate diagnosis and effective treatment management plan in India. Innovative medical technologies are essential for patient care. This is also true for prevention of various diseases related to hygiene, communicable diseases, addiction related diseases like lung cancer, oral cancer, liver cirrhosis etc. In future the scope of technology applications like data mining techniques based systems in the healthcare system in India will really make dramatic changes at every level. Today the internet is a gateway to world knowledge as well as a massive platform for national media and documentation. This will help a lot in future in implementation of data mining techniques

V. CONCLUSION

For effective utilization of data mining in health organizations there is a need of enhance and secure health data sharing among different parties. Some propriety limitations such as contractual relationships among researcher and health care organization are mandatory to overcome the security issues. There is also a need of standardized approach for constructing the data warehouse. In recent years due to enhancement of internet facility a huge datasets (text and non-text form) are also available on website. So, there is also an essential need of effective data mining techniques for analyzing this data to uncover hidden information.

REFERENCES

[10] Leigh turner,First world health Care at Third World Prices’: Gobllization,Biotechs And Medical Tourism