# Introducing Health Care Analytics for Cancer Treatment

*Nitesh Dugar[1], Surendra Yadav[2]*

[1]Research scholar, JECRC University , Jaipur
nitraj.dugar@gmail.com
[2]Associate Professor, JECRC University, Jaipur
surendra.yadav@jecrc.edu.in

*Abstract--Hospital supervision or healthcare administration is the field with reference to management, leadership and administration of hospitals, health care systems and hospital networks. Healthcare business these days generates huge amounts of complicated information concerning with patients, medical devices, electronic patient records and sickness designation, hospital resources etc. The huge amounts of data's information and knowledge may be a key resource to be processed and analyzed for knowledge extraction that permits support for cost-savings and higher cognitive process. Data mining brings a collection of tools and techniques that may be applied to the present processed data's information to getting hidden patterns that offer healthcare professionals an additional or extra source of information and knowledge for better decisions for the administration.*

*Keywords—Hospital, Data Mining, Healthcare, Extraction, Hidden Patterns*

## 1. INTRODUCTION

The healthcare business is one of the largest and fastest growing industry in the world, it consume around 10 percent of the Gross Domestic Product (GDP)of most of the developed or developing nation, healthcare contribute a huge part for a country's economy. Indian Healthcare industry provides new and existing players with a unique and special opportunity to achieve innovative research and profits. Healthcare in India also awarded as 'polio Free' country by World Health Organization (WHO). According to McKinsey & Company in the next decade ,consumer awareness and demand for better facilities will increase and in India healthcare will become second largest service sector employer.

Healthcare information is diverse in scope and huge in content and its volume is so vast that traditional / routine analytical methods reveal very little of the possible conclusions. Modern data mining techniques can be applied to this data to extract otherwise hidden / unknown facets of knowledge which may be of vital importance to therapeutic, commercial and preventive aspects of healthcare.

The new innovation in healthcare Data Mining is 'Big Data Analytics' revolution. The health care Big data means electronic health data sets which are so complex and large that they are nearly impossible to manage with traditional hardware machine and software tools and techniques. For the health care data scientist, there is huge amount of data is available to discover and understand the pattern and trends inside the data hence big data analytics has potential to improve healthcare services such as care, life and cost reduction.

This paper involves the discussion on various data mining algorithms used in health care analytics.

## 2. DATA MINING

Data Mining explores the hidden patterns or information from data warehouse. This knowledge i.e. information extracted from huge dataset is present in human understandable form. Therefore Data mining is also Known as KDD (Knowledge Discovery).Data Mining tools and techniques helps to predict future trends and behavior of a business which helps in making proactive and knowledge driven decisions. Data Mining techniques and tools can answer the complex business questions which were difficult and time consuming to resolve.

The foundation of Data mining techniques are the result of a long process of analytical research and come up with a product development. Data mining application initially used by business community because it supports:

- Massive data collection
- Powerful multiprocessor computer
- Data mining algorithms.

Later on it comes in Healthcare domain for HMIS (Health care Management Information System) and detection of fraudulent cases. Health care and commercial databases are growing at unpredictable rate. To handle these huge data sets we need matured data mining algorithms which can come out with older statistical methods.

## 3. HEALTHCARE ANALYTICS USING DATA MINING

Generally health care data is available in flat files, relational databases or in advanced database systems such as images which is collected from different data sources such as: OPD, Laboratory information system, Operation theatre module, Radiotherapy module, Chemotherapy module, Blood bank and Drug store

A. **Nature of Healthcare data:** The healthcare data is very specific. To mine healthcare data all information need to be changed into numeric values. . The methods for this task are described in medical textbooks which usually come from certain range and it is beyond the scope of this thesis. With

the help of numeric data it is easy to apply mining algorithm to extract knowledge out of data.

B. **Patient Data Set:** A patient dataset is a collection of patient information which is collected from the different sources as mentioned above. In this thesis we have collected data source from GCRI (Gujarat Cancer and Research Center) and prepared dataset for different regions of Gujarat and Surrounding states and applied Association Rule, Clustering and Classification .This helps us to determine cancer site and morphological pattern among various patients and to develop suggestive management information system to improve cancer treatment.

C. **Preliminary Analysis of dataset:** The main objectives of our preliminary analysis is to identify the hospital and government needs, perform economic, technical analysis and green analytics, perform cost benefit analysis and create suggestive management information system. There should be enough expertise available for database query and data mining algorithms and software's for doing analysis.

Before performing analysis, the following questions emerge:

- How much time should be spent on creating suggestive management information system? As such, there are no rules or formulas available to decide on this. However, dataset size, complexity end-use, contractual obligation is few parameters on which it should be decided.

- How to bifurcate the datasets? There is around 64 GB of unstructured data present in the database for different kind of diseases. Therefore we have structured the data and created dataset for cancer patients to determine cancer site and morphology pattern among patients from various regions of Gujarat and surrounding states.

Other major question that arises is which type of data mining algorithms we should apply? Firstly we have studied

different types of algorithms we can use on healthcare data and then we applied as Association Rule, Clustering and Classification on our dataset

## 4. MEDICAL DATA SELECTION & PREPARATION

Medical data selection & preparation (MDP) is very crucial, important and very time consuming process for Data Mining. It took around 45 to 55 % of time to prepare data set for the whole Data Mining process. Health care data selection and processing aims at establishing a common data warehouse for Data Mining algorithms and information sharing. MDSP is connected to various data sources such Laboratories, Operation Theater, Blood Bank, Drug Store, Therapy Modules which holds valuable data, such as patient diseases information, his work flow and past prescriptions.

The Medical Data selection process select's data from different sources as mentioned above. And all the relevant data is updated at centralized data warehouse according to the patient details. Then, MDS process select whether the data is useful or not and then passed it to Research Department. After collecting data from various sources, the next phase is data processing. Data processing plays an important role in significant role in entire data mining process as it insures quality and accuracy of data.

Medical data Preparation involves examining or logging the data and verifying the data for accuracy and exactness. This is because some selected data may be missing or in different formats. In this stage all necessary information's are extracted from the data selection process to apply for further Data Mining process like the age of patient should be "81" but recorded as "18" so this comes under the human error. Data cleaning can reduce the missing, inconsistency and noisy data that affect the results.

## 5. CANCER TREATMENTS USING DECISION SUPPORT SYSTEM

Cancer is the social health problem in the world and one of the leading causes of death. According to WHO (world Health Organization) approx. 40 percent of all cancer type are treatable but it is a chronic diseases. And in most of the cancer risk factors are known and most common are smoking, alcohol and some pernicious habits, adipose (excessive weight).New medical technologies are emerging into the market to identify cancer at initial stage to improve life length of patient with cancer. Most of the patient consult to experts having symptoms of the last stage of a diseases, which limits to patient care i.e. treatment .In order to diagnosis it in early stage a set of action should be taken place. Therefore to diagnose it in early stage decision support systems are used.

The main objective of decision support system is to support medical expert with additional information and knowledge which helps them to make decision whether a patient needs endoscopy or not. Decision support system(DSS) contains two main modules: Data Mining module and decision support module's in medical science primarily done in two areas, firstly in lower level such as hospital & patient management, diagnosis and treatment, keeping records, finance management and inventory management. Second is the higher level decision making that helps in giving hospitals a competitive environment.

A DSS product name called as Isabel, a web based system contains electronic patient record. It contains two parts: Isabel Diagnosis Reminder System (IDRS) and Isabel Knowledge Mobilizing System (IKMS) .IDRS gives list to experts for given set of clinical feature that are based on human systems. IKMS has dictionary of thousands of diagnosis categories which helps experts to do concept search rather than keyword search.

To design DSS two factors are involved stake holder involvement and type of decision he want to make. While designing the health care DSS, there should be organized design like procedure for decision making, strategic planning and structure according to government should be kept in mind. The effectiveness of DSS depends on methodology used to design the system. We can divide it into three approaches:

- Clinical algorithms
- Heuristics approaches

- Mathematical approaches

Some experts believe that medical knowledge contains various patterns was more effective.

Neural networks methodologies are also use to design DSS but use is limited to laboratory and medical image applications. Data scientist revealed that neural network systems provide the planning for patient care, length of patient stay and the mortality rate. However to implement DSS by this methodology extensive research and resources are required .Another very popular approach is data mining techniques which helps to identify rules and patterns about various problems. Data mining DSS is built based on the data and it is effective in cost reduction and improving quality of care. Another approach is Strategic Healthcare Decision Support Services (SHDS) which is in synergy between data mining and knowledge management.

Issues seen in DSS in healthcare are to improve quality, safety and efficiency of patient care and hospital. The main challenge is cost of the DSS investment is fuzzy and there is always a debate about what should be included or excluded while calculating cost and benefit for system. Apart from cost second major issue is quality and speed of decision making .As the expertise (nurses and doctor) have to deal with various complex diagnoses work, it become time consuming for them to adopt with a new technology or a system. That is why, DSS not accepted widely due to time complexity and constraint but it is perceived effective and efficient to use.

## 6. APPROACHES AND ALGORITHMS OF DATA MINING IN CONTEXT OF HEALTH CARE

A data mining algorithm is a set of numerical calculations that creates a data mining model from data. To create data mining model, the data is primarily provided to algorithm for analyzing and then it look for specific type of patterns and trends.

The data mining model creates from data can take following forms:

- A set of cluster to show how the dataset are related in database.
- A decision tree which will predict the outcome and it also define how different criteria affects the outcomes.
- A mathematical model
- A set of pattern which describe how product is cluster together in a transaction, and the probability of purchasing product simultaneously.

Choosing the correct algorithm for a specific analytical task is challenge. While using different algorithms for same task, each algorithm will produce different result.

Selection of algorithm by its type:

- **Classification algorithm:**
  It will predict more than one discrete variables based on the other attributes in the data.
- **Regression algorithms** :
  It will predict more than one continuous variable based on the other attributes in the data.
- **Segmentation algorithms** :
  It will divide the data into cluster or group of similar kind of object.
- **Association algorithms** :
  It finds similar relationship between different attribute in a dataset. Mostly it is used to create association rules which are used in market basket analysis.
- **Sequence analysis algorithms**: It will summarize the frequently episodes of data.

Data mining allows you to build n number of model on a single mining structure, therefore within a single mining solution we can use decision model, clustering algorithm etc to get multiple views of data.

| Example of task | Algorithm to use |
|---|---|
| **Predicting a discrete value** | Decision Trees Algorithm |
| | Naive Bayes Algorithm |
| | Clustering Algorithm |
| Categorization of | Neural Network Algorithm |

| | |
|---|---|
| patient according to its outcomes and explores related factors.<br><br>Calculating the probability of server failure within the next 4 months. | |
| **Predicting a continuous attribute**<br><br>Predict next year sales<br><br>Calculating or generating a risk factor or score on a given demographics. | Decision Trees Algorithm<br>Time Series Algorithm<br>Linear Regression Algorithm |
| **Sequence Prediction**<br><br>Capturing and analyzing the sequential activities of outpatient visit, to formulate common activities. | **Sequence** Clustering Algorithm |
| **Searching similar items in transactions:**<br><br>Use of market basket analysis.<br><br>Suggesting additional product to buyer for purchase | Association Algorithm<br>Decision Trees Algorithm |

TABLE1. SELECTING ALGORITHM BY TASK

## 7. CONCLUSION

Health issues are arising fast in this era. Scientists are running behind the technologies to find best to best solutions for diseases. Data mining is one of the best solutions to treat patients with the help of past experiences and knowledge extracted from the data collected in previous years. Data mining has a great impact in the field of health care. Heath care industries are improving their outputs with the use of various techniques and equipments developed by medical scientists.