Data Mining Techniques in Medicine for Analyzing Cardiac Patient Data

KEYWORDS

ByAizaj M. Kazi	Dr. Sudhakar D. Bhoite
Asst. Professor,Dept. Of Computer Studies, SIBER, Kolhapur, MS, India,	Asso. Professor, Dept. Of M.Phil (Com & Mgt.),SIBER, Kolhapur, MS, India

Introduction

In the last couple of decades the disease burden across the world has shifted from communicable diseases to Non-Communicable Diseases (NCDs). Moreover, a majority of people suffering from NCDs reside in the developing countries. The nations, not having completely deal with the course of communicable diseases yet, are now facing the additional burden of NCDs. The already inadequate and stretched healthcare system in different countries has meant that the mortality from NCDs is high.

When we talk about human lives, the heart is heart of human body, without which human is as good as a dead body. It provides the filtered blood to various organs in the body, for its proper functioning. The diagnosis of diseases with in human body is the significant and complex task in medicine. The detection of heart diseases from considering various factors or symptoms is a multi-layered issue, which is not free from false presumptions, often accompanied by unpredictable effects. Thus an attempt to utilize knowledge and experience of the specialist's and clinically screening the data of patients collected and stored in databases, which further facilitate the diagnosis process. The healthcare industry gathers enormous amount of heart disease data that regrettably, are not "mined" to discover useful information for effective decision making by healthcare practitioners in terms of knowing symptoms, performing diagnosis, for better prescription so as to serve better treatment to the patient.

Cardiovascular Diseases (CVDs) were once thought was impacting the human body of any age and gender. While changing lifestyle, unhealthy eating habits and declining physical activities are the key reasons for high incidences of CVD's, the issues of access and affordability account for higher mortality among the population. These diseases impact not only on the wellbeing of human, but can also hold back the economic growth of the country, due to increased healthcare, expenditure and diminished productivity. India is projected to be high rates of CVDs and diabetes in coming years of duration.

0.1. Healthcare Industry In India

A Healthcare service in India has undergone vast changes over the last few decades which encompass the entire nation. The India is expected to supersede China by 2030 in terms of population expansion. Hence, it becomes one of the prime duties of the state to improve the nutrition status and the standard of living of the people, together with improving public health to large extent.

The rapid increase in health care industry of our country is one of country's largest sector, both in terms of revenue and employment. It has been estimated that the healthcare industry of India will grow by 40 billion. The continuous increase in the population growth of India is considered as one of the principal reason for the growth in the healthcare industry of India. The rise in the communicable as well as chronic degenerative diseases, has contributed to the rise in the healthcare sector of India.

Between 2008 and 2030, the global population is projected to grow by 20%, from 6.7 billion to 8.1 billion people. The crude death rate is expected to remain more or less stable at around 8.4 deaths per thousand. However, a major shift is currently underway in the overall diseases burden on the world. In 2008, five out of the top ten causes for mortality worldwide, other than injuries, were non-communicable diseases; this will go up to seven out of ten by the year 2030. By then, about 76% of the deaths in the world will be due to non-communicable diseases (NCDs).

Top Ten causes of mortality

2008	
Cardiovascular diseases	Cardiovascular diseases
Cancers	Cancers
Chronic Respiratory diseases	Chronic Respiratory diseases
Respiratory infections	Respiratory infections
Perinatal Conditions	Diabetes Mellitus
Diarrhoeal diseases	Digestive diseases
Digestive diseases	Perinatal Conditions
HIV/AIDS	Neuropsychiatric
Tuberculosis	Genitourinary diseases
Neuropsychiatric conditions	HIV/AIDS

Non-Communicable Diseases (NCD) Communicable Diseases

Source: Global Burden of Diseases 2004. Projected Deaths 2030, Baseline Scenario World Health Organization, 2008 Number of deaths in thousand

India Ranks 37th in the world for the heart diseases. It is estimated that by 2020 cardiovascular disease will be the cause of over 40 percent deaths in India as compared to 24 percent in 1990. Globally, it causes 17.3 million deaths annually. With over 3 million deaths owing to cardiovascular diseases every year, India is set to be the 'heart disease capital of the world' in few years, said doctors on the eve of World Heart Day (September 29, 2014)

Research Design
2.1 INTRODUCTION

Since from the decades the Health care industry flourished in region of western Maharashtra of Sangli and Kolhapur, Miraj is well known for the hospital industry not in India

but in other country of Gulf countries such as Oman, Qatar KSA, Dubai. Sangli is the area which consist of the Mission Hospital (Walness Hospital, Miraj) Civil Hospital and Bharati Hospital Walnesswadi, Sangli providing various type of health care facility from Nephrology, Urology, Neurology, Orthopedic, Dental and Cardiology and specialization in healthcare area since the thousands of the patients from various areas from the Maharashtra and some part of Karnataka are visiting the for the health related problems.

The data mining techniques can be applied to such a data for finding the fruit full information for finding the health related problems. This research is primarily intended to investigate the role of data mining techniques in the medicine of cardiac patients

2.2 SIGNIFICANCE OF THE STUDY
The present research focus upon the efficient approach of Data mining for the extraction of significant information for pattern recognition of the symptoms, diagnosis of problems of the cardiac patients and derivation of treatment. Data mining assist end users to extract useful information from voluminous database and discovers the meaningful information. Data mining helps to discover the knowledge which is previously unknown. Various areas such as Marketing, Customer Relationship, Engineering, Logistic, Government, Financial, Banking, Retailing, Manufacturing, Production, Import, Export, Mobile Computing, and Communication are increasingly using data mining, for gaining knowledge out of heaps of data.

The medicinal field also rapidly moving towards standardization in terms of recording (digitizing) of patient symptoms, diagnosis, and treatment, surgical methods, medical tests and relationship among them, with online systems and increasingly applying data mining to streamline the operations.

2.3 STATEMENT OF THE RESEARCH PROBLEM
It is mentioned, In the recent survey carried out by WHO (World Health Organization in 2011) that maximum number of Cardiac diseases causes increasing numbers of deaths worldwide as compared to other diseases. As the current study aims to find out where the problems lies for the cardiac, and what are its symptoms, treatment given by the practitioner and in what age group does it lies and in which gender does it found. In this study researcher has made an attempt to evaluate the various symptoms for cardiac problems.

The study is carried out by conducting a sample survey of doctors and patients to collect the primary data. Further this has enabled the researcher to analyze and diagnose the issues and challenges in the topic under study. Hence the statement of the problem is entitled as *"Role of Data Mining in Medicine of Cardiac Patients in Sangli City"*

2.4 OBJECTIVES OF THE STUDY
The objectives of this research study were as follows:
1. To study the issues, challenges in cardiac related problems and, how to overcome the challenges and issues.
2. To study the symptoms of cardiac patients.
3. To apply suitable data mining techniques for mining symptoms and treatment related data.
4. To suggest some remedial solutions for betterment of society, if needed.

2.5 SCOPE OF THE STUDY
The geographical scope of study is restricted to Sangli city

only. The present study focuses upon cardiac problems, symptoms and relevant treatment. The data is collected from physicians and patients by keeping in view major issues related to cardiac problems. The topical scope will be to test the hypothesis and perform the analysis on surveyed data. The technical scope is to make use of suitable data mining tools, for performing clustering, association on the collected data, which has helped the researcher to draw relevant inference, and to fulfill the objectives set forth while carrying out this study.

2.6 Research Design
2.6.1 Population Under Study
The population for study includes data from the 20 cardiac patients collected from the 15 hospital out of the 30 hospital by using random sampling method doctors which are doing the practice in the cardiology the list of doctors are collected from the IMA (Indian Medical Association) list of Sangli city

RESEARCH METHODOLOGY AND DATA SOURCES
Data was collocated from list of registered patents was procured and were present and willing to participate in survey was considered for the study the patient Data was entered into excel sheet, for further processing, by applying data mining tools as WEKA data is also be tabulating the primary data for performing data analysis and interpretation

2.6.2 Primary Data
Structured questionnaire scheduled were used for collecting data while carrying out this study. A comprehensive questionnaire with 36 various questions under six sections was been prepared and a pilot study was carried out after the pilot study suitable changes were made to the questionnaire and trial study was conducted. Each section was designed to collect data about certain topic of health related problems. Data is collected from the patients through questionnaires from any age group, gender.

2.6.3 Data Collection Techniques
Questionnaire were prepared for all respondents both in Marathi and English language which were filled by the respondents.

2.6.4 Sampling Methodology Adopted:
The population of study is 20 respondents of the patients from 15 doctors practicing in the cardiology. Data was collocated from list of registered patents was procured and who were present and willing to participate in survey was considered for the study.

Stage 1: Selection of Hospitals
There are totally 30 hospitals with doctor practicing in cardiology in Sangli city. In the first stage, out of 30 hospitals 15 hospital are selected in the sample for the study of cardiac patient. The sample hospitals were selected randomly using lottery method by writing the names of doctor on piece paper. 50% hospitals from the jurisdiction of Sangli city has been selected in the sample.

Stage 2: Selection of Patients (Respondents):
Purposive sampling method is used in second stage of sampling for selecting the Patient from each selected Hospital. 50% of respondent from each Hospital is selected for the study. There are 640 patients visits the hospital OPD, out of which 300 are selected for the study. Data was collocated from list of registered patents was procured and who were present and willing to participate in survey was

considered for the study.

DATA ANALYSIS

Analysis is done by applying data mining tool ie: WEKA. The data is analyzed in different dimensions with its category and summarized by identifying the relationship(s). As data mining is the process of discovering the knowledge finding correlation, pattern among the surveyed data.

Data analyzed with the data mining tool WEKA which helps in faster analysis and interpretation of data along with different parameters

Table 5.1 Age group wise cluster for blood pressure

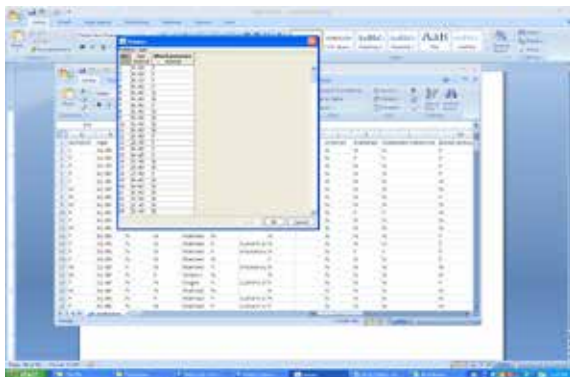| Age Group	Yes	No	Total
01-20	8	3	11
	72.70%	27.30%	100.00%
21-40	60	13	73
	82.20%	17.80%	100.00%
41-60	116	14	130
	89.20%	10.80%	100.00%
61-80	66	16	82
	80.50%	19.50%	100.00%
Above 80	4		4
	100.00%		100.00%
Total	254	46	300
	84.70%	15.30%	100.00%

Figure 5.1 Data of the above table

Figure 5.2 Age wise blood pressure clusters

Table : Age wise Cholesterol level clusters

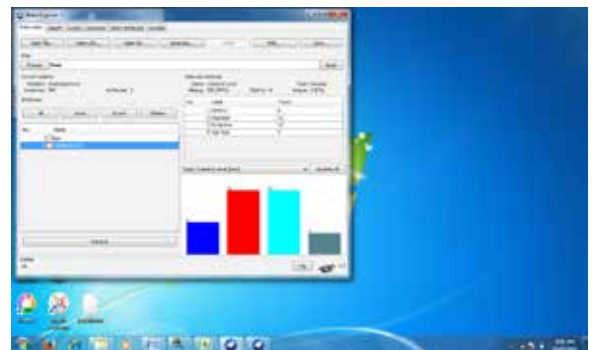| Age Group	150-200 (Desirable)	200-240 (Borderline)	240-400 (High Risk)	I Don't Know	Total
21-40	2	2			4
	50.00%	50.00%			100.00%
41-60	5	2	2	6	15
	33.30%	13.30%	13.30%	40.00%	100.00%
61-80	5	8	2		15
	33.30%	53.30%	13.30%		100.00%
Total	12	12	4	6	34
	35.30%	35.30%	11.80%	17.60%	100.00%

Fig : 5.5 WEKA preprocess on the csv file format of cholesterol
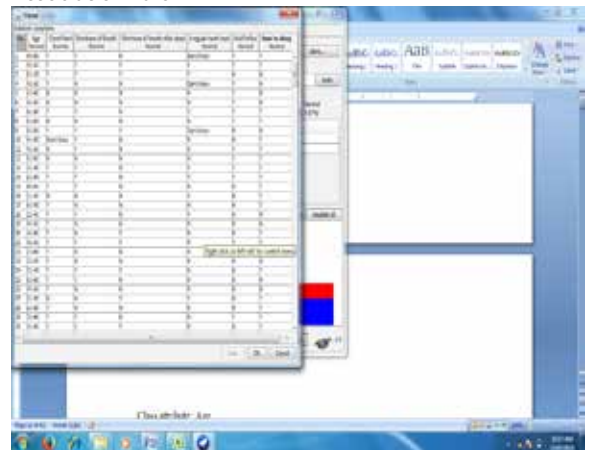
Association rule

Fig : 5.7 WEKA preprocess on the csv file format of on symptoms

The output of the *Apriori* algorithm by applying association rule on various symptoms.

== Run information ===

Scheme: weka. associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

Relation: symptoms

Instances: 300

Attributes: 7

 Age
Chest Paint
Shortness of Breath
Shortness of breath after sleep
Irregular heart beat
Acid Reflux
Snore in sleep

Apriori
Minimum support: 0.2 (60 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 16
Generated sets of large itemsets:
Size of set of large itemsets L(1): 14
Size of set of large itemsets L(2): 42
Size of set of large itemsets L(3): 35
Size of set of large itemsets L(4): 7

Rules found:
1. Shortness of Breath=Y Shortness of breath after sleep=Y Acid Reflux=N 71 ==> Chest Paint=Y 67 conf:(0.94)
2. Shortness of Breath=Y Shortness of breath after sleep=Y 106 ==> Chest Paint=Y 99   conf:(0.93)
3. Shortness of breath after sleep=Y Irregular heart beat=N 70 ==> Chest Paint=Y 65   conf:(0.93)
4. Shortness of Breath=Y Acid Reflux=Y Snor in sleep=Y 70 ==> Chest Paint=Y 65   conf:(0.93)
5. Shortness of breath after sleep=Y Acid Reflux=N 79 ==> Chest Paint=Y 73   conf:(0.92)
6. Shortness of Breath=Y Acid Reflux=Y 86 ==> Chest Paint=Y 79   conf:(0.92)
7. Chest Paint=Y Shortness of breath after sleep=Y Acid Reflux=N 73 ==> Shortness of Breath=Y 67 conf:(0.92)
8. Shortness of Breath=Y Irregular heart beat=N Acid Reflux=N 73 ==> Chest Paint=Y 67   conf:(0.92)
9. Shortness of Breath=Y Irregular heart beat=N 130 ==> Chest Paint=Y 119   conf:(0.92)
10. Age=41-60 Shortness of Breath=Y 91 ==> Chest Paint=Y 83   conf:(0.91)

One of the most popular data mining approaches is to find frequent item sets from a transaction dataset and derive association rules. Finding frequent item sets (itemsets with frequency larger than or equal to a user specified minimum support) is not trivial because of its combinatorial explosion. Once frequent item sets are obtained, it is straightforward to generate association rules with confidence larger than or equal to a user specified minimum confidence. Apriori is a seminal algorithm for finding frequent itemsets using candidate generation

It is characterized as a level-wise complete search algorithm using anti-monotonicity of item sets, "if an item set is not frequent, any of its superset is never frequent". By convention, Apriori assumes that items within a transaction or item set are sorted in lexicographic order.

Conclusions:
The Present research study is in progress, while in the pivoted study some data set used from internet and the data is analyzed and interpreted. The above data reveals that majority of the human deaths takes place due to CVD problems and next to that is due the Cancer problem. Thus it further calls for a detailed study to understand the causes of CVD. Also The study helps in constraining and summarizing algorithms of data mining used in the field of medicine for prediction are discussed. The focus is was on using different algorithms and combinations of several target attributes for intelligent and effective heart attack prediction using data mining. For predictions of heart attack, significantly 26 attributes were considered Association Rules approach.

REFERENCE 1. AshaRajkumar,Mrs. G. Sophia Reena"Diagnosis Of Heart Disease Using Data mining Algorithm",Vol. 10 Issue 10 Ver. 1.0 September 2010Global Journal of Computer Science and TechnologyGJCST ClassificationJ.3 | 2. Elma Kolce (Çela), NekiFrasheri"A Literature Review of Data Mining Techniques Used in Healthcare Databases", ICT Innovations 2012 Web Proceedings - Poster Session ISSN 1857-7288 | 3. Farhan IsmaeelDakheel, RaoofSmko, K. Negrat, AbdelsalamAlmarimi, "Using Data Mining Techniques for Finding Cardiac Outlier Patients", World Academy of Science, Engineering and Technology 49 2011 | 4. JyotiSoni, Ujma Ansari, Dipesh Sharma, SunitaSoni, "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction" International Journal of Computer Applications (0975 – 8887) Volume 17– No.8, March 2011 | 5. N. Aditya Sundar, P. PushpaLatha, M. Rama Chandra, "Performance Analysis of Classification Data Mining Techniques Over Heart Disease Data Base", International Journal of Engineering Science & Advanced Technology ISSN: 2250–3676 [IJESAT] Volume-2, Issue-3, 470 – 478 | 6. NidhiBhatla, Kiran Jyoti, "An Analysis of Heart Disease Prediction using Different Data Mining Techniques", International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 8, October - 2012 ISSN: 2278-0181 | 7. N. S. Nithya, S. Sarumathi, Dr. K. DuraiSwamy"Assessment of the Risk Factors of Heart Attack Using Frequent Feature Selection Metho" International Journal of Communications and Engineering Volume 01– No.1, Issue: 01 March2012