Abstract—Data mining aims at discovering novel, interesting and useful information from databases. Conventionally, the data is analyzed manually the drawback of which is many hidden and potentially useful information may not be understood by the analyst. Presently, many organizations including modern hospitals are collecting huge amounts of data which requires an automated way to extract useful knowledge. Therefore medical domain is a major area for applying data mining. Through data mining, interesting knowledge can be extracted and applied to increase the working efficiency. In this write-up the salient features of data mining in medicine are discussed.

Keywords—data mining, medical data mining, features of data mining, uniqueness, early warning.

I. INTRODUCTION

The Healthcare industry is among the most information intensive industries. Medical information, knowledge and data keep growing on a daily basis. It has been estimated that an acute care hospital may generate five terabytes of data a year [1]. Medical informatics plays a very important role in the use of clinical data. In such discoveries pattern recognition is important for the diagnosis of new diseases and the study of different patterns found when classification of data takes place. It is known that “Discovery of HIV infection and Hepatitis type C were inspired by analysis of clinical courses unexpected by experts on immunology and hepatology, respectively”[2].

Computer assisted information retrieval may help support quality decision making and to avoid human error. Although human decision-making is often optimal, it is poor when there are huge amounts of data to be classified. Also efficiency and accuracy of decisions will decrease when humans are put into stress and immense work. This lead to the use of data mining in medical informatics, the database that is found in the hospitals, namely, the hospital information systems (HIS) containing massive amounts of information which includes patients information, data from laboratories which keeps on growing year after year. With the help of data mining methods, useful patterns of information can be found within the data, which will be utilized for further research and evaluation of reports.

Healthcare is research intensive field and the largest consumer of public funds. With the emergence of computers and new algorithms, health care has seen an increase of computer tools and could no longer ignore these emerging tools. This resulted in uniting of healthcare and computing to form health informatics. Health informatics “deals with biomedical information, data, and knowledge--their storage, retrieval, and optimal use for problem solving and decision making”[3]. This is a highly interdisciplinary subject where fields in medicine, engineering, statistics, computer science and many more come together to form a single field. With the help of smart algorithms and machine intelligence we can provide the quality of healthcare by having, problem solving and decision-making systems. Information systems can help in supporting clinical care in addition to helping administrative tasks. Thus the physicians will have more time to spend with the patients rather than filling up manual forms.

II. DATA MINING

Data mining (the analysis step of the "Knowledge Discovery in Databases” process, or KDD), an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data preprocessing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization and online updating.

Data Mining has been an area of emerging technology over the past decade, picking up speed and growing rapidly as various industries realize the potential of such practice. Since the inception of this relatively new area, work owes and ultimately understandable patterns in data. Knowledge discovery is therefore holding increasing relevance to medical areas, and this is frequently being acknowledged within the research literature. For example, Kudyba [6] states...
that ‘strong disease management programs depend on data mining techniques’, and he also refers specifically to asthma. Given the vast amounts of medical data available today, data mining and artificial intelligence techniques appear to offer the best method of knowledge discovery. Data mining may allow for the identification of natural groups of patients given inputs such as symptoms, and further classify or predict derivatives from the given data.

III. FEATURES OF DATA MINING

A. Uniqueness:- Human medical data are most rewarding and difficult of all biological data to mine and analyze. Humans are the most closely watched species on earth. Human subjects can provide observations that cannot easily be gained from animal studies, such as visual and auditory sensations, the perception of pain, discomfort, hallucinations, and recollection of possibly relevant prior traumas and exposures. Most animal studies are short-term, and therefore cannot track long-term disease processes of medical interest, such as pneumoepithelialis or atherosclerosis. With human data, there is no issue of having to extrapolate animal observations to the human species. The ‘uniqueness’ of medical data mining, as described by Cios and Moore[7] may be categorically defined into four groups: a) Heterogeneity of medical data, b) Ethical, legal, and social issues, c) Statistical philosophy and d) Special status of medicine. These groups present problems and barriers, perhaps due to traditional training methods that thwart the bridging of the knowledge gap.

a. Heterogeneity of medical data

The major areas of heterogeneity of medical data are Volume and complexity of medical data (raw medical data are huge, voluminous and heterogeneous). These are collected through interviews with the patients, various images from diagnostic techniques and clinical impression, diagnosis). Importance of physician’s interpretation (the physician’s interpretation of diagnostic results, even specialists from same discipline differ in describing a patient’s condition hence it becomes difficult to mine),[8,9]. Sensitivity and specificity analysis (precision in nearly all diagnosis and treatments in medicine is subject to errors, there is need to distinguish between a test and diagnosis in medicine. The medical condition of the patient is characterized by various tests and based on these tests ailment is diagnosed. Both tests and diagnosis are subject to sensitivity and specificity analysis.). Poor mathematical characterization (medical data cannot be put into formulas equations and models because the conceptual structure of medicine consists of word descriptions and images) and Canonical form (in biomedicine even elementary concepts have no canonical forms hence consistency in expression becomes difficult).[10]

b. Ethical, legal, and social issues

Since the medical data are collected on human beings, it involves ethical, legal and social issues that are designed to prevent the abuse of patients and misuse of their data. These issues may pertain to data ownership, fear of law suits, privacy and security of human data, expected benefits and administrative issues.

c. Statistical philosophy

Statistical philosophy of medical data is different from non-medical data. Medical data is collected to provide benefit the individual patients. Sometimes the patient who consents to be involved in the research projects may not get any benefit but the data collected from such patients is narrowly focused and is regulated by legal and ethical considerations. [11]

d. Special status of medicine

Medical science enjoys special status in day to day life. Medical care is a necessity; it gives life to the patient and hope for a fruitful life. Every patient expects recovery after going to the medicine person and demands ethics, sympathy and care but no one realizes the risk which medicine person face. Legal cases against doctors are on the increase. Everybody wants to enjoy the benefits of research but no one is ready to contribute into it. The collected medical data when published is used for social causes without harming the dignity of the patients. Though medical science enjoys special status yet it has to face many hindrances. It is still debatable as to what questions to be asked from patients, what tests to be performed on the patients and what conclusions may not be drawn. That is why experiments on animals are conducted and results from these experiments are considered reliable.

B. Early Warning:

Data mining on medical data has great potential to improve the treatment quality of hospitals and increase the survival rate of patients. Clinical study has found that 4–17% of patients undergo cardiopulmonary or respiratory arrest while in the hospital [12]. Early detection and intervention are essential to preventing these serious, often life-threatening events. Indeed, early detection and treatment of patients with sepsis has already shown promising results, resulting in significantly lower mortality rates [13]. Mao et al proposed an early warning system (EWS) based on data mining technology which is designed to identify the science of clinical deterioration and provide early warning for serious clinical events. The early warning system of Mao et al. is designed to provide reliable early alarms for patients at the general hospital wards (GHWs). EWS automatically identifies patients at risk of clinical deterioration based on their existing electronic medical record. The main task of EWS is a challenging classification problem on high dimensional stream data with irregular, multi-scale data gaps, measurement errors, outliers, and class imbalance. They have proposed a novel data mining framework for analyzing such medical data streams. The framework addresses the above challenges and represents a practical approach for early prediction and prevention based on data that would realistically be available at GHWs. [14]
Ngan et al. have discovered medical knowledge by learning Bayesian networks and rules. Evolutionary computation is used as the search algorithm. The Bayesian networks can provide an overall structure of the relationships among the attributes. The rules can capture detailed and interesting patterns in the database. The system is applied to real-life medical databases for limb fracture and scoliosis. The knowledge discovered provides insights to and allows better understanding of these two medical domains [15].

IV. RELATED WORK

Alon et al.[16] (1999) analyzed gene expressions in 40 tumor and 22 normal colon tissue samples with an affymetrix oligonucleotide array complementary to more than 6500 human genes. They applied two way clustering algorithm both genes and tissues that revealed broad coherent patterns that suggest a high degree of organization underlying gene expression in these tissues. Clustering also separated cancerous from noncancerous tissue and cell lines from in vivo tissues on the basis of subtle distributed patterns of genes even when expression of individual genes varied only slightly between the tissues. Two-way clustering thus may be of use both in classifying genes into functional groups and in classifying tissues based on gene expression.

Sung-Bae Cho and Hong-HeeWon (2006) [17] have used representative gene vectors that are highly discriminatory for cancer classes and extracted multiple significant gene subsets based on those representative vectors respectively. Also, an ensemble of neural networks learned from the multiple significant gene subsets is proposed to classify a sample into one of several cancer classes. The performance of the proposed method is systematically evaluated using three different cancer types: Leukemia, colon, and B-cell lymphoma.

Hualong Yu, et al. (2012)[18] have reported skewed gene selection algorithm that introduces a weighted metric into the gene selection procedure. The extracted genes are paired as decision rules to distinguish both classes, with these decision rules then integrated into an ensemble learning framework by majority voting to recognize test examples; thus avoiding tedious data normalization and classifier construction. The mining and integrating of a few reliable decision rules gave higher or at least comparable classification performance than many traditional class imbalance learning algorithms on four benchmark imbalanced cancer gene expression datasets.

Khaya (2012) [19] have summarized various review and technical articles on breast cancer diagnosis and prognosis. She also focused on current research being carried using the data mining techniques to enhance the breast cancer diagnosis and prognosis. Breast cancer diagnosis is distinguishing of benign from malignant breast lumps and breast cancer prognosis predicts when breast cancer is to recur in patients that have had their cancer excised.

Ada and Kaur (2013) [20] used data mining techniques such as Data Processing, Feature Extraction, Classification using neural network and SVMs to classify digital X-ray film into two categories i.e. normal and abnormal with respect to lung cancer. Normal or negative ones are those characterizing a healthy person, abnormal or positive ones include lung cancer patient. Thus the data mining techniques helped in early detection of lung cancer.

V. DISCUSSION

Health informatics is an emerging field. It is especially important as it deals with collection, organization, storage of health related data. With the growing number of patient and health care requirements, having an automated system will be better in organizing, retrieving and classifying of medical data. Physicians can input the patient data through electronic health forms and can run a decision support system on the data input to have an opinion about the patient’s health and the care required. An example in the advances in health informatics can be the diagnosis of a patient’s health by a doctor practicing in another part of the world. Thus healthcare organizations can share information regarding a patient which will cut costs for communication and at the same time will be more efficient in providing care to the patient. Data Mining techniques have been extensively used in cancer research, where it has helped in characterizing the patients in different categories. Data Mining on medical data has great potential in developing an early warning systems which will help in initiating the proper treatment of life threatening diseases.

VI. CONCLUSION

In short data mining is different from all fields because the data are heterogeneous and special ethical, legal and social constraints are applied to private medical information. Methods of medical data mining must address heterogeneity of data sources, data structures, and the pervasiveness of missing values for both technical and social reasons. For all its perils, medical data mining can also be the most rewarding. For an appropriately formulated medical question, finding an answer could mean extending a life, or giving comfort to an ill person. These potential rewards more than compensate for the many extraordinary difficulties along the pathway to success. Using medical data mining it becomes possible to improve the treatment quality of hospitals and increase the survival rate of patients.

REFERENCES


14. Mao Yi, Chen Yixin, Hackman Gregory, Chen Minmin, Lu Chenyang, Kollef Marin and Bailey Thomas C., Medical data mining for early detection warning in general hospitals”.


