Smartcare as an Application of Data Mining

Jyoti G. Daga[1] Neha N. Gaonkar[2] Kalpana D. Kajale[3]
[1,2,3]B.E Student
[1,2,3]Department of Computer Engineering
[1,2,3]Atharva College of Engineering Mumbai, India

Abstract— Data mining as one of the many constituents of health care has been used extensively in many organizations around the world as an efficient technique of finding correlations or patterns in large relational databases which results into more pragmatic health information. In healthcare, data mining is becoming increasingly popular and essential. Data mining applications will be an asset to all parties involved in health care industry. The huge amounts of data generated by healthcare transactions are too complex and mammoth to be processed and analyzed by traditional methods. Data mining provides the method to transform huge amount of data into useful information for decision making. This paper looks at the data mining applications in healthcare, it discusses data mining and its applications in major areas of health informatics.
A major objective is to test data mining tools in medical and healthcare applications to develop a tool that can help make perfect and correct decisions. A brief summarization of various data mining algorithms used for classification, clustering, and association as well as their respective pros and cons is also presented. A discussion of the technologies available to enable the estimation of healthcare costs (including length of hospital stay), disease diagnosis and prognosis is offered along with a discussion of the use of data mining to discover such relationships as those between existing conditions and a disease, relationships among diseases, and relationships among drugs. The main objective is to mine the data available to predict doctors in particular area, age wise disease frequency, medicines recommended by most doctors.

Key words: Data mining, healthcare application, Data mining algorithms

I. INTRODUCTION

The Health industry is among the most information specific fields. Medical information keeps growing on an everyday basis. It has been estimated that a major hospital may generate five terabytes of data a year. The ability to use this data to extract useful information for efficient healthcare is of prime importance. In such developments pattern recognition is important for the diagnosis of new diseases and the study of existing ones. Although man's ability to make decisions is often optimal, it is vulnerable to errors when there is an immense amount of data to be classified. Automatic classification is done based on patterns present in the data. This automatic classification is called as data mining.

Data mining can be defined as the process of finding previously undiscovered patterns and trends in databases and using that information to create predictive models. It could also be defined as the process of data selection and exploration and building models using massive data stores to uncover previously unexplored patterns. Data mining is an analytic process designed to explore massive amounts of data in search of relevant patterns and/or systematic relationships between variables, and then to test the findings by applying the detected patterns to new subsets of data. Huge amounts of data generated by healthcare transactions are too complex and mammoth to be processed and analysed by traditional ways hence calls for technology to simplify management of that data. Data mining can improve decision making by discovering patterns and trends in extensive amounts of complex data. Such analysis has become intensely important as financial pressures have increased the need for healthcare organizations to make decisions supported by the analysis of clinical and financial data. Healthcare organizations that perform data mining are in an exceedingly higher position to meet their future needs; data can be a great aid to healthcare organizations, but it has to be first remodelled into information.

Yet another factor motivating the employment of data mining applications in healthcare is the realization that data mining can generate information that is very beneficial to all parties involved in the healthcare industry. For example, healthcare providers can gain relevant assistance in making decisions. Data mining applications also can benefit healthcare providers such as hospitals, clinics, physicians by identifying effective treatments and best practices. The aims of quality healthcare services are:
providing safe healthcare treatments

- using scientific medical knowledge to provide healthcare services to everyone
- providing various healthcare treatments based on the patient's wants, symptoms and preferences
- minimizing the delay time in providing medical treatment.
- Health outcomes (e.g. mortality, disability, well-being)

II. BACKGROUND HISTORY

Though the term data mining was introduced in Nineties, the idea of data mining has its roots since many years. Data mining reached its current stage by going through many phases of analysis and studies. This growth began once business information began to get stored on computers. The process continued with advancements in computer technology coupled with information

storage, processing power, new software system, algorithms etc. Moreover, in today's competitive world of information, all are attempting to make the most effective use of their information to make their businesses profitable and prosperous.

Next evolutionary step in data processing happened throughout Nineteen Eighties with the introduction of relational databases and structured query language. Thus, data became accessible at record level dynamically. Next was the introduction of knowledge reposition happened throughout Nineties. Online analytic processing and multidimensional databases contributed to the expansion of data warehousing. Now, the rising technique is data processing. If every step of evolution is analyzed, it's crystal clear that every step is constructed upon the previous step.

During Sixties information was simply information wherever the case is currently entirely different. Business data has been reworked to business information that is powerful enough to answer several complicated business queries and even to foresee the long run of business. Anyhow, methoding technologies are researching development process for several years and 3 completely different areas contributed to the expansion of knowledge mining in its current form. Those areas are statistics, artificial intelligence and machine learning.

Statistics has been conducive greatly to business intelligence for the previous few decades. Anyhow, statistics wasn't thriving in responding to complex business questions of today's competitive business world. Still, the ideas of statistics additionally cope with data and relationships among them. These ideas are the building blocks of advanced data mining techniques.

III. REVIEW OF LITERATURE

[3] This research involves the development of personalized electronic health record system for monitoring patients with chronic conditions, that

a) allow for relevant data to be entered by the patient,
b) make relevant data available to patient's care provider, at real-time and at doctor's visi
c) will generate reports and graphs for the data and
d) will provide secure storage of the data.

The Personalized Electronic Health Record System for Monitoring Patients with Chronic Disease (PEHRSMPCD) is designed to permit tracking and monitoring of the symptoms of patients with chronic disease and provide healthcare professionals with data on patients' lifestyle changes, medication (drug) changes, diet changes and symptom changes. A Personalized Electronic Health Record System application would allow for relevant data to be entered by the patient to facilitate the tracking and monitoring of symptoms of patients with chronic illness.

Effective management of Hospital resource data processing provides support for constructing a model for managing the hospital resources that is a vital task in aid. Data mining is a process that finds helpful patterns from great deal of knowledge the method of extracting antecedently unknown, comprehendible and unjust data from giant databases and transforming it to form crucial business selections. This data processing definition has business flavour and is for business environments.

International Journal of Recent Development in Engineering and Technology Website: (ISSN 2347-6435(Online) Volume 3, Issue 3, September 2014) 25 Numerous studies highlighted that data processing techniques facilitate the knowledge holder to investigate and find out unexpected relationship among their data that are successively useful for creating call.

Hallick has advised that data processing techniques are useful to produce the knowledge to patient concerning numerous diseases and their bar.

Salim, Suzan, Daniel, Dina, Anael compared and contrasted causes, symptoms, and courses of treatments, data mining can deliver an analysis of which courses of action prove effective. Other data mining applications related to treatments include associating the various side-effects of treatment, collating common symptoms to aid diagnosis, determining the most effective drug compounds for treating sub-populations that respond differently from the mainstream population to certain drugs and determining proactive steps that can reduce the risk of afflicting future needs of individuals to improve their level of satisfaction

Insights gained from data mining can influence cost, revenue and operating efficiency while maintaining a high level of care. Healthcare organizations that perform data mining are better positioned to meet their long term needs; data can be a great asset to healthcare organizations, but they have to be first transformed into information. Yet another factor motivating the use of data mining applications in healthcare is the realization that data mining can generate information that is very useful to all parties involved in the healthcare industry.

This knowledge can lead to better diagnosis and treatment for future patients. Data mining and knowledge discovery is the name often used to refer to a very interdisciplinary field, which consists of using methods of several research areas to extract knowledge from real world data sets. There is a distinction between the terms data mining and knowledge discovery; the term data mining refers to the core steps of a broader process called knowledge discovery in database.

Gallatas, Zitos and Fillia [4] applied well known feature selection and data mining algorithms such as forward selection and Naïve Bayes respectively, to determine patient satisfaction, which is an important indicator of quality of care in hospital settings.

Using this framework with their data, they concluded that important factors for patient satisfaction is the medical unit specialty, the patient's length of stay, the technological infrastructure of the hospital, whether the hospital is collaborating with a university, the opinion of the nurses about the quality of provided care and the number of beds of the hospital.

Finally, they found out that the patients' level of education, their perceived care from the healthcare professionals, their views about the hospital environment as well as their opinions about the process followed when they left the hospital may be important

factors which influence patient satisfaction. Patient Satisfaction Data concerning the patient treatment in healthcare is useful in understanding the factors that affect patient satisfaction and how personal experience of the patients is related to them.

IV. IMPLEMENTATION DETAILS

Data mining consists of varied strategies. Different strategies serve different purposes, every methodology has its benefits and drawbacks. Data mining tasks may be divided into descriptive and predictive. Whereas descriptive tasks have a goal on finding an individual's interpreted forms and associations, when reviewing the information and the whole construction of the model, prediction tasks tend to predict an outcome of interest.

Though the goals of description and prediction tasks might overlap, the main distinction is that the predictive ones need that data include a special variable of response. The response may be categorical or numeric.

Additionally classifying data mining as classification and regression. [8] Predictive tasks make it practical to predict the worth of a variable based on different existing data. Descriptive tasks, on the opposite hand, mix the information in a very certain method. the most predictive and descriptive data processing tasks may be classified as following:

A. Classification And Regression

These tasks are predictive and they involve the creation of models to predict target, or dependent variable from the set of explained or independent variables. Classification is the method of finding a function that permits the classification of information in one amongst many categories. For classification tasks, the target variable typically includes a small number of separate values, whereas with the regression tasks, target variable is continuous.

B. Association Rule

association rule analysis is descriptive data processing task which has determinative patterns, or associations, between components in data sets. Associations are depicted in the form of rules, or implications.

C. Cluster Analysis

descriptive data processing task with the goal to group similar objects within the same cluster and different|completely different} ones within the different clusters. Method of grouping determines groups of data that are similar, however different than other information. During this method variables are known by which the best grouping is being realized.

D. Text Mining

most of the accessible information is in the form of unstructured or partly structured text, and it's completely different from standard information that are fully structured. Text is unstructured if there's no previously determined format, or structure in information. Text is partly structured if there's a structure linked with the parts of information. Whereas text mining tasks typically constitute classification, clustering and association rule data processing classes, it's the most effective to look at them individually, because unstructured text demands a particular thought. in particular, method for illustration of textual information is crucial.

E. Link Analysis

kind of network analysis that examines the associations between objects. Link classification provides class of an object, not simply based on its features, however also on connections in which it takes part, and features of objects connected with certain path. Example of link analysis in medicine is task of predicting illness type based on people's characteristics or predicting age of people on the basis of malady they're infected with and based on age of individuals they have been in touch with. Link analysis may be utilized in order to grasp where patients go to receive the health care treatment and to spot the elements or resources in commission that must be self-addressed. this is a data mining type that has population tracking throughout their movement from point to point within the system. This analysis requires population segmentation so the analysis will target proportion of the population. in order for the link analysis to be pragmatic, all the patient's data should be stored in databases (personal data, dates and time of visits, doctors that treated the patient, doctors that gave referrals, patient's previous diseases)

Upon completion of the data analysis, all results are displayed in a very clear manner, typically within the kind of tables or diagrams that will be 2 dimensional or 3 dimensional. Programs even enable the user to vary any of the variables, and also the impact of its modification is shown in real time on the diagram.

Data mining algorithms are required in virtually each step in KDD method starting from domain understanding to data analysis. Each data processing technique serves a diverse purpose depending on the modelling objective. the two most typical modelling objectives are classification and prediction. Classification models predict categorical labels (discrete, unordered) whereas prediction models predict continuous-valued functions.

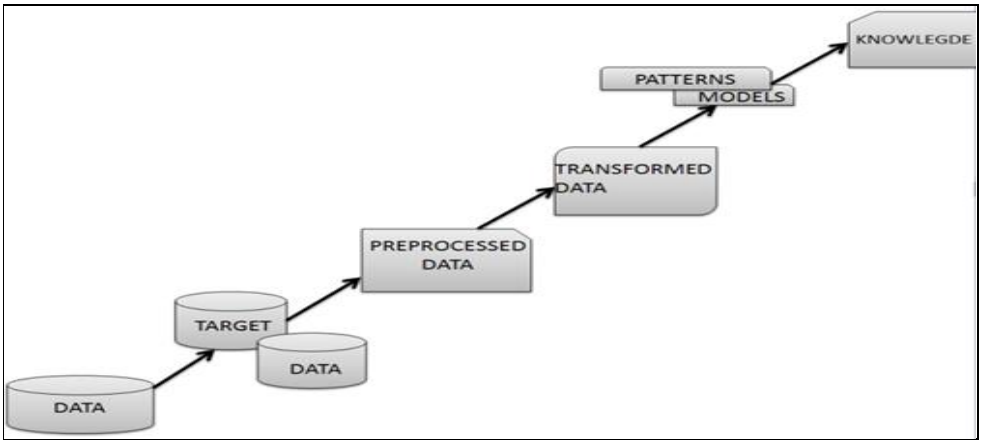

Fig. 1: Knowledge data discovery (KDD)

Decision Trees and Neural Networks use classification algorithms while Regression, Association Rules and Clustering use prediction algorithms. Here are some of the data mining algorithms which are successfully used in healthcare.

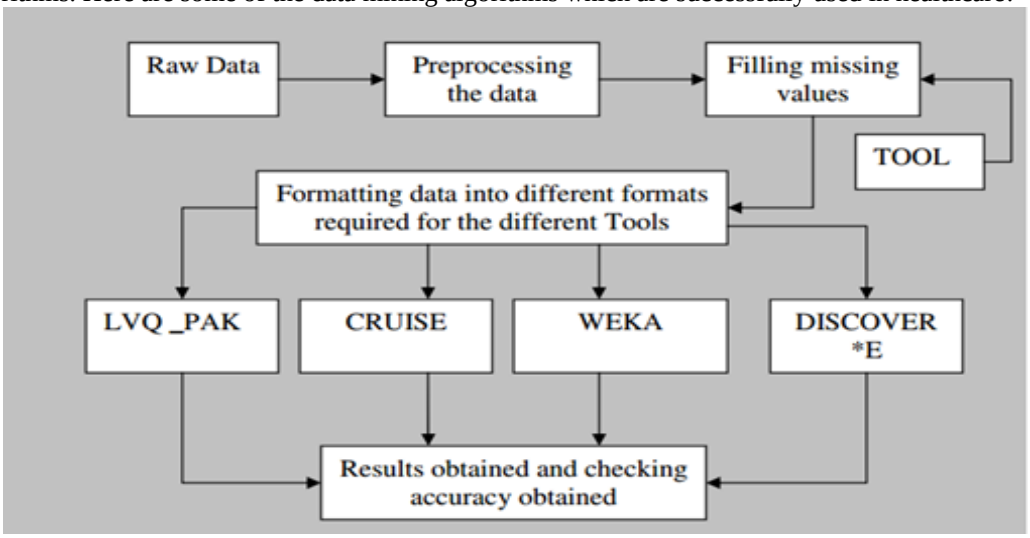

Fig. 2: Tools for Data mining

1) Naïve Bayes

The Naïve Bayes is a simple probabilistic classifier. Naïve Bayes is based on the assumption of mutual independency of attributes. The algorithm works on the assumption, that variables provided to the classifier are independent. The probabilities applied in the Naïve Bayes algorithm are calculated using Bayes Rule. The Bayesian formalism is a way of representation of uncertainties what is essential during diagnosis, prediction of patients' prognosis and treatment selection [7]. It is possible to present the interactions among variables using Bayesian networks. These networks are often understood as cause-and-effect relationships. The application of a Bayesian network in medicine was presented for instance in diagnosis and antibiotic treatment of pneumonia by P. Lucas [6]. In addition, the Naïve Bayes method's performance was tested against a colorectal cancer in which the authors enhanced the effectiveness of this method

2) Decision Trees

Decision trees are one of the most regularly used techniques of data analysis. Decision trees are easy to visualize and understand and resistant to noise in data. Generally, decision trees are used to classify records to a proper class. Besides, they are applicable in both regression and associations tasks. In the medical field decision trees specify the sequence of attributes values and a decision that is based on these attributes Decision Trees can also handle continuous data but they must be converted to categorical data. The decision trees are effectively applied in medicine for instance in prostate cancer classification.[8] Here C4.5 algorithm was used. The article presents a study carried out to create a decision tree model to describe how women in Taiwan make a decision whether or not to have a hysterectomy. The qualitative study was conducted and a tree model was built. This method, based on the Galwin's methodology, had accuracy of 90%.

3) Neural networks

Artificial neural networks are analytical techniques that are formed on the basis of superior learning processes in the human brain. As the human brain is capable to, after the learning process, draw assumptions based on previous observations, neural networks are also capable to predict changes and events in the system after the process of learning. Neural networks are groups of connected input/output units where each connection has its own weight. The learning process is performed by balancing the net on the basis of relations that exist between elements in the examples. Based on the importance of cause and effect between certain data, stronger or weaker connections between "neurons" are being formed. Network formed in this manner is ready for

the unknown data and it will react based on previously acquired knowledge. Artificial neural networks are ideal for multiprocessor systems, where a large number of operations are performed in parallel.

One of the key advantages of Artificial Neural Networks is their high performance. The core functions of Artificial Neural Networksis prediction. The disadvantage of this method is its complexity and difficulty in understanding the predictions. Their effectiveness and usefulness was proven in medicine. The successful implementation of the neural networks was in the development of novel antidepressants. The notable success is the application of a neural network in coronary artery disease and processing of EEG signals.

4) Genetic algorithms

Genetic algorithms are based on the principle of genetic modification, mutation and natural selection. These are algorithmic optimization strategies inspired by the principles observed in natural evolution. The genetic algorithm creates a number of random solutions to the problem. All these solutions may not be good, a group of solutions can be skipped entirely, and it can come down to the overlapping solutions. Poor solutions are discarded, and the good ones retained. Good solutions are then being hybridized, and then the whole process is repeated. Finally, similar to the process of natural selection, only the best solutions remain. So, from the set of potential solutions to the problems that compete with each other, the best solutions are chosen and combined with each other in order to obtain a universal solution from the set of solutions that will become better and better, similar to the process of evolution of organisms. Genetic algorithms are used in data mining to formulate hypotheses about the dependencies between variables in the form of association rules or other internal formalism. The disadvantage of this method is that it requires an huge amount of processing power and it is too slow for trivial issues.

5) Apriori Algorithm

Apriori algorithm for association is proposed by R.Agarwal et al., in 1994. It finds out the relationships among item sets using two inputs-support and confidence. These two inputs help to discriminate the frequent and infrequent item sets. The research work filtered out those item from transaction database that are not satisfy some given criteria such as frequent item set satisfy the minimum support and confidence constraint. This algorithm is based on the principle that if an item does not fulfils minimum support constraint or not frequent then its descendants are also not frequent so remove this item from the transaction database because this item does not contribute in the construction of association rules. Unlike classification and clustering, efficiency is the evaluation factor of association mining. Various methods are used to improve the efficiency of Apriori algorithm such as Hash table, transaction reduction, partitioning etc.

Patil et al. [5] used apriori algorithm for generating association rule. Using these rules they classify the patients suffering from type-2 diabetes. In this research, authors proposed an approach for discretizing the attributes having continuous value using equal width binning interval which is selected on the basis of medical expert's opinion. Another research work analyzes the medical bill using apriori algorithm. Abdullah et al.,[5] proposed some modification in existing Apriori algorithm and then utilize its effectiveness in constructed useful information in medical bill. Ilayaraja et al., [5] also used Apriori algorithm to discover frequent diseases in medical data. This study proposed a method for detecting the occurrence of diseases using Apriori algorithm in particular geographical locations at particular period of time.

V. CONCLUSION

In the paper we have proposed the integration of data mining into healthcare applications. It is our belief that the paper will be a contribution to the data mining and healthcare literature and practice. We have given a holistic perspective of the various tools of data mining that can be used in healthcare outlining applications of data mining in healthcare. It also is hoped that this paper can help all parties involved in healthcare to reap the benefits of healthcare data mining.

ACKNOWLEDGMENT

We would like to thank our project guide Prof. Neha Singh for her enormous co-operation and guidance. We have no words to express our gratitude for a person who wholeheartedly supported the project and gave freely of her valuable time while making this project. All the inputs given by her have found a place in the project. The technical guidance provided by her was more than useful and made the project successful. She has always been a source of inspiration for us. It was memorable experience learning under such a highly innovative, enthusiastic and hardworking teacher. We are also thankful to our Principal Dr. S.P. Kallurkar, our HOD Professor Mahendra Patil, our Project co-ordinator Professor Deepali Maste and all the staff members of the Computers department who have provided us various facilities and guided us to develop a very good project idea. Finally, we would also like to thank teachers of our college and friends who guided and helped us while working on the project.

REFERENCES

[1] Data Mining Techniques and Application in a Healthcare. Battu Vani1 B.Balakrishna M.Tech Scholar1, Assistant Prof.2, TITS, Hyderabad, 2004
[2] Overview application of data mining in health care, Salim, Suzan, Daniel, Anael
[3] Personalized Electronic Health Record System for Monitoring Patients with Chronic Disease Imran A. Khan, Member, IEEE
[4] Application of Data Mining Techniques to Determine Patient Satisfaction Georgios Galatas, Dimitrios Zikos, Fillia Makedon Heracleia lab, CSE.
[5] A survey on Data Mining approaches for Healthcare. Divya Tomar and Sonali Agarwal

[6] Applying Data Mining Techniques to a Health Insurance Information System Marisa Viveros, John Nearhos, Micheal Rothman
[7] Study and Analysis of Data mining Algorithms for Healthcare Decision Support System Monali Dey, Siddharth Swarup Rautaray
[8] Data Mining In Healthcare: A Survey of Techniques and algorithms with its limitation and challenges Prakash Mahindrakar1, Dr. M. Hanumanthappa2
[9] Prediction and Decision Making in Health Care using Data Mining Boris Milovic*, Milan Milovic**
[10] Approaches to Partition Medical Data using Clustering Algorithms, P.Kalyani Research Scholar of Mother Teresa Women's University, Kodikanal.