ISSN: 2321-7782 (Online)
Volume 4, Issue 2, February 2016
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

An Overview of Data Mining Applications in Healthcare

Ravleen Singh[1]
Lecturer G.D.C. Poonch - India

Dr. Tariq Hussain Sheikh[2]
Lecturer G.D.C.Mendhar - India

Abstract: ***There is a wealth of data available within the healthcare systems. However, there is a lack of effective psychoanalysis tools to discover hidden relationships and trends in data. Without data mining it is difficult to understand clearly the full potential of data collected within healthcare organization as data under analysis is exceptionally large, highly dimensional, distributed and not definite. The objective of this study is to explore new and emerging areas of data mining techniques used in healthcare management. This paper aims to make a detailed study report of different types of data mining applications in the healthcare area and to minimize the complexity of the study of the healthcare data understanding.***

Keywords: Data Mining.

I. INTRODUCTION

Data mining provides the methodology and technology to transform massive amount of data into useful information for decision making. It is defined as the process of data selection and exploration and building models using vast data stores to uncover unknown patterns [1].The investigative objective of data mining is to organize the data, text and images of huge data into knowledge based or information using dispensation of computers. Data mining algorithms applied in healthcare sector play a significant role in prediction and diagnosis of the diseases [2]. The large numbers of data mining applications are found in the medical related areas such as Medical device industry, Pharmaceutical Industry and Hospital Management. The research objective of data mining in Health Care System is to generate an automated tool to perceive, ascertain voluminous data and organize into useable information i.e. Knowledge discovery data (KDD).

Fig 1.Data Mining Architecture

The knowledge discovery is an interactive process, consisting by developing an understanding of the application domain, selecting and creating a data set, preprocessing, data transformation [3]. Data mining tools answer the question that traditionally was a time consuming and too complex to resolve [4]. They prepare databases for finding predictive information. Data mining tasks are Association Rule, Patterns, Classification, Prediction and Clustering. Most common modeling objectives are classification and prediction [5]. The branch of computer science which is more actively and efficiently involved in medical sciences is Artificial Intelligence. Any computer program that helps experts in making healthcare decision comes under the domain of healthcare decision support system.

Fig 2.Data Mining Life Cycle

II. DATA MINING METHODOLOGY

Data mining procedure performs following maneuvers proposed by the Cross-Industry Standard Process [6],[7]. The main motto acknowledged by business understanding step lead towards the emergence success of data mining project. The data understanding step furtherance into preliminary data collection, data description, data exploration and verification of data quality. After the identification of data in data preparation step, it should be selected, clean, built in the desired format and formatted. The modeling step diagnosis the actual data analysis into modeling stage. There are different data mining software's viz. cluster analysis, discriminant analysis, regression analysis, neural analysis network, decision tree, link analysis and association analysis. The suitable model is applied to data types. Further the evaluation step compares the data model resulting from any data mining model by using any common standard techniques such as lift chart, profit chart and diagnostic classification chart. Finally deployment resulted which is related to the actual execution and operationalization of data mining model.

III. PROCESS OF KNOWLEDGE DISCOVERY AND DATA MINING

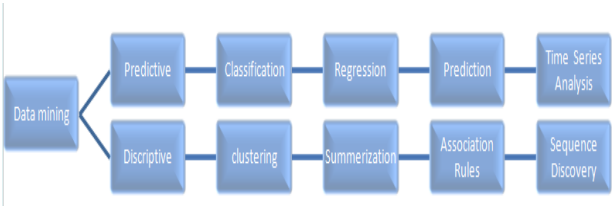


Fig 2.Data mining model and task

C1.Predictive data modeling is an important data mining task to determine future data states on the basis of past and current values. Predictions may be made on the basis of regression, time series analysis, or some other approaches.

Time series analysis is to identify the value or attributes over a time period usually at evenly spaced time intervals i.e. The attribute value can be generated on a daily or hourly basis depending upon the state of the ailing which is used to foresee the future analysis.

Regression is a method to map target data using some known type of function. It deals with estimation of an output value based on input values.

Classification – is the task of generalizing known structure to apply to new data. Classification classifies a data item into one of several predefined classes. A set of classification rules is generated from the classification model, based on the features of the data in the training set, which can be used to classify future data and develop a better understanding of each class in the database.

C2. Visualization techniques are useful methods of discovering patterns in a medical data set. Scatter diagrams in a Cartesian plane of two interesting medical attributes can be used to identify interesting subsets of medical data sets. For

example, for heart patients interesting subsets can be found with respect to blood sugar (fasting). Once interesting subsets are obtained, we may use other data mining techniques on these subsets to discover further knowledge.

Association rule (Dependency modeling) –Searches for relationships between variables. For example a supermarket might gather data on customer purchasing habits.

Clustering – is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.

Summarization – providing a more compact representation of the data set, including visualization and report generation

Link analysis – Form of network analysis that examines the associations between objects Link. Classification provides category of an object, not just based on its features, but also on connections in which it takes part, and features of objects connected with certain path [8].

IV. DATA MINING TECHNIQUES

There are two basic design of data mining: hypothesis testing and knowledge discovery [**9**]. Hypothesis testing is a top-down approach that is used when a confirmation or a rejection of an already defined hypothesis is needed. The knowledge discovery is a bottom-up approach and it is used when we want to find something that we do not know from searching available data. In Data mining the most important technique which is used is Knowledge Discovery in Database(KDD).KDD has different steps like Data cleaning, Data integration, Data selection, Data transformation, Data mining, Pattern evaluation, Knowledge presentation etc. The different types of techniques used in Data mining project include Decision tree, Bayesian networks, Naive bayes, Neural networks etc.

D1. 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. C4.5 is used in classification problems and it is the widely used algorithm for building DT. It is suitable for real world problems as it deals with numeric attributes and missing values.

D2. Naive Bayse- this is a modest probabilistic classifier, which is based on an assumption about mutual independency of attributes. The probability which is applied in the Naïve Bayes algorithm are calculated according to the Bayes Rule, The naive Bayes classifier's attractiveness is in its simplicity, computational efficiency, and good classification performance. The naïve Bayes classifier requires a very large number of records to obtain good outcomes.

D3. Neural Networks -There are 3 layers in neural networks: input layer, hidden layer, output layer. Hidden layer is the products of the input layer. The condition between neurons has weights which are assigned to them. Their values are calculated with the use of back Propagation algorithm. In hidden layers there are some nonlinear features added to the network. The out layer may have more than one output node which predict the different diseases.

D4. Genetic algorithms -are based on the standard of genetic modification, mutation and natural selection. These are algorithmic optimization strategies motivated by the principles observed in natural evolution [10]. 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.

D5. Nearest neighbor -method is a method that is also used for data classification. Unlike other techniques, there is no learning process to create a model. The data used for learning is in fact a model. When the new data shows up, the algorithm analyzes all the data in the database to find a subset of instances that are the best fit and based on that it is able to predict the

outcome. The study [11] conducted on the application of nearest neighbor method on standard data set to detect efficiency in the diagnosis of heart diseases, produced the results that application of this method achieved an accuracy of 97.4% which is a higher percentage than any other published study on the same set of data. More details on data mining techniques can be found in Berry and Linoff [12]

Intelligent decisions are similar to human decisions but are automated decisions. Classification and prediction in machine learning are among the techniques that can produce intelligent decision [13]. So are researches still in progress. Decision tree models are best suited for data mining. They are inexpensive to construct, easy to interpret, easy to integrate with database system and they have comparable or better accuracy in many applications. There are many Decision tree algorithms such as HUNTS algorithm (one of the earliest algorithm), CART, ID3, C4.5 (a later version ID3 algorithm), SLIQ, SPRINT [14]. Artificial neural networks (ANN) provide a powerful tool to help doctors to analyze, model and make sense of complex clinical data across a broad range of medical applications. A neural network has been successfully applied to various areas of medicine, such as diagnostic aides, medicine, biochemical analysis, and image analysis and drug development [15].

V. Data Mining Evolving Applications

There is vast potential for data mining applications in healthcare. Some of them can be grouped as under:

E.1 Infection Control in Hospital

Two million patients each year in the United States are affected by Nosocomial infections. Computer-assisted surveillance research has focused on identifying high-risk patients, expert systems, and possible cases and detecting deviations in the occurrence of predefined events [16]. The system uses association rules on culture and patient care data obtained from the laboratory information management systems and generates monthly patterns that are reviewed by an expert in infection control. An early warning of the global spread of SARS virus is an example of the usefulness of syndrome's systems based on data mining [17].

E2. High-Risk Patients Identification

American Health ways provides diabetes disease management services to hospitals and health plans designed to enhance the quality and lower the cost of treatment of individuals with diabetes. A robust data mining and model building solution identifies patients who are trending toward a high-risk condition. WEKA 3.6 is used as the data mining tool to implement the Algorithms. The J48 classifier performs classification with 81.8% accuracy in predicting the HIV status [18].

TABLE 1. DATA MINING APPLICATION IN HEALTHCARE

S.NO	Types of Disease	Data Mining Tool	Technique	Algorithms	Traitional Methods	Accuracy Level(%) of Data Mining
1	Heart Diseases	ODND NCC2	Classification	Naïve	Probability	60
2	Cancer	WEKA	Classification	Rule DesisionTable		97.77
3	HIV/AIDS	WEKA 3.6	Classification,Association Rule Mining	J48	Statistics	81.8
4	Blood Bank Sector	WEKA	Classification	J48		89.9
5	Brain Cancer	K-MEANS Clustring	Clustring	MAFIA		85
6	Tuberculosis	WEKA	Naïve Bayes classifier	KNN	Probability ,Statistics	78
7	Diabetic Mellitus	ANN	Classification	C4.5 Algorithm	Neural Network	82.6
8	Kidney Dialysis	RST	Classification	Decision making	Statistics	75.97
9	Dengu	SPSS MODEL		C5.0	Statistics	80
10	Hepatitis C	SNP	Information Gain	Decision making		73.2

E3.Treatment effectiveness

Data mining applications can be developed to evaluate the effectiveness of medical treatments [19]. By comparing the outcomes of patient groups treated with different drug regimens for the same disease or condition can be compared to determine which treatments work best and are most cost-effective. Data mining could be particularly useful in medicine when there is no dispositive evidence favoring a particular treatment option Based on patients' profile, history, physical examination, diagnosis and utilizing previous treatment patterns, new treatment plans can be effectively suggested for Examples: Onset, treatment and management of depression [20]. Treatment Decision Support Tool for Patients with Uterine Fibroids[21]. HianChyeKoh and Gerald Tan works on data mining and its applications with major areas like Treatment effectiveness, Management of healthcare, Detection of fraud and abuse, Customer relationship management[22].

E4.Healthcare management

To aid healthcare management, data mining applications can be developed to well identify and track chronic disease states and high-risk patients, design appropriate interventions, and reduce the number of hospital admissions and claims. Sierra Health Services has used data mining comprehensively to identify areas for quality improvements, including treatment guidelines, disease management groups, and cost management [23].

E5.Healthcare Resource Management

Effectively manage the resource allocation by identifying high risk areas and predicting the need and usage of various resources. If the inpatient length of stay (LOS) can be predicted efficiently, the planning and management of hospital resources can be greatly enhanced. Neural network system is used to predict the disposition in children presenting to the emergency room with bronchiolitis.

E6.Customer relationship management

Customer interactions may occur through call centers, physicians' offices, billing departments, inpatient settings, and ambulatory care settings. The principles of applying of data mining for customer relationship management in the other industries are also applicable to the healthcare industry. In many cases prediction of purchasing and usage behavior can help to provide proactive initiatives to reduce the overall cost and increase customer satisfaction.

E7.Pharmaceutical Industry

The pharmaceutical firms manage their inventories and to develop new product and services. Pharmaceutical companies can benefit from healthcare CRM and data mining by tracking which physicians prescribe which drugs and for what purposes. Pharmaceutical companies can decide whom to target, show what the least expensive or most effective treatment is plan for an ailment [24].

E8.Fraud and abuse

Data mining applications that attempt to detect fraud and abuse often establish norms and then identify unusual or abnormal patterns of claims by physicians, laboratories, clinics, or others. Among other things, these applications can highlight inappropriate prescriptions or referrals and fraudulent insurance and medical claims. A method based on naive Bayes that effectively combines the advantages of boosting and the explanatory power of the weight of evidence scoring framework was presented in [25] Furthermore, the classification algorithm C4.5 was applied for fraud/abuse detection by using the discovered temporal patterns as predictive features. A data mining framework that uses the concept of clinical pathways (or integrated care pathways) was utilized for detecting unknown fraud and abusive cases in a real-world data set gathered from the National Health Insurance (NHI) program in Taiwan [26].

E9.Human Talent Predictions

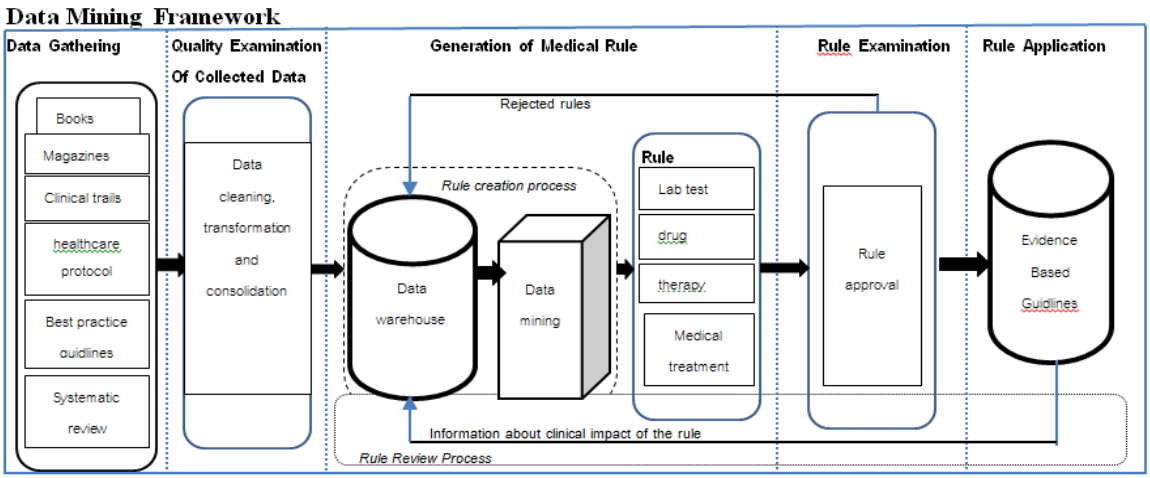
We can discover an employee performance using classification and prediction techniques in DM. Since the construction of decision trees does not require any expert knowledge or parameter setting, they remain popular and are considered for exploratory knowledge discovery. Still the technique which is otherwise known as the 'divide-and-conquer' rule is undergoing researches.

E10.Talent Forecasting

Association rules are used to associate employee's profiles to the most appropriate program or job and then associate the employee attitude with performance. And the predictions used on classification to find out the percentage of accuracy in employee performance, behavior and attitude, analyzing, forecasting and identifying the best profile for different employees

E11.System Biology

Biological databases contain a wide variety of data types, often with rich relational structure. Accordingly multi-relational data mining techniques are frequently applied to biological data [27]. Systems biology is at least as demanding as, and perhaps more demanding than, the genomic challenge that has fired international science and gained public attention.



VI. LIMITATION IN DATA MINING

Data mining applications can greatly benefit the healthcare industry but they were some limitations as well [28]. Healthcare data mining can be limited by the accessibility of data, because the raw inputs for data mining often exist in different settings and systems [29], such as administration, clinics, laboratories and more. Data from heterogeneous sources present the challenges. Secondly, the missing, corrupted, inconsistent or non-standardized data such as pieces of information recorded in different formats in different data sources is a big challenge. Missing attribute values can impact the assessment of whether a particular combination of attribute-value pairs is significant within a dataset [30]. Cleaning data from noise and outliers and handling missing values, and then finding the right subset of data, prepares them for successful data mining [31]. Thirdly, there may be ethical, legal and social issues, such as data ownership and privacy issues, related to healthcare data. Fourthly, the successful application of data mining requires knowledge of the domain areas as well as in data mining methodology and tools otherwise the user may not be able to avoid the pitfalls of data mining [32]. Fifthly a sufficiently exhaustive mining of data will certainly yield patterns of some kind that are a product of random instabilities [33] especially for large data sets with many variables. Hence, many interesting or significant patterns and relationships found in data mining may not be useful. Finally, Data mining projects can fail for a variety of reasons, such as lack of management support, unrealistic user expectations, poor project management, inadequate data mining expertise, and many more.

VII. Conclusion and Future Work

The key to successful data mining is to first define the business or clinical problem to be solved. New knowledge is not discovered by the algorithms, but by the user. Today, there have been many efforts with the goal of successful application of data mining in the healthcare institutions. Primary potential of this technique lies in the possibility for research of hidden patterns in data sets in healthcare domain. These patterns can be used for clinical diagnosis. However, available raw medical data are widely distributed, different and voluminous by nature. These data must be collected and stored in data warehouses in organized forms, and they can be integrated in order to form hospital information system. There are extra research challenges for the integration of such an ample data for information system. As healthcare data are not limited to just quantitative data, such as physicians' notes or clinical records it is also useful to look into how digital diagnostic images can be brought into healthcare data mining applications. Some progress has been made in these areas [34], [35] with the future development of information communication technologies; data mining will achieve its full potential in the discovery of knowledge hidden in the medical data which is concerned with research edifice.

References

1. Tan, P., Steinbach, M. and Kumar, V. Introduction to Data Mining, Addison-Wesley, Boston, 2006.
2. grawal, R. and Srikant, R., 1994. Fast algorithms for mining association rules. In Proceedings of the 20th International Conference on Very Large Databases (VLDB ″ 94), Santiago, Chile.
3. Agrawal, R. and Shim, K., 1996. Developintightly coupled data mining applications on a relational database system. In Proceedings of the 2nd International Conference on Knowledge Discovery in Databases and Data Mining (KDD ″ 96), Portland, Oregon, USA.
4. Berry, M. J. A., & Linnof, G, Data mining Techniques, New York: Wiley, (1997).
5. Milley, A. (2000). Healthcare and data mining. *Health Management Technology*, 21(8), 44-47.
6. CRoss Industry Standard Process for Data Mining[Online] Available:http://www.crisp-dm.org, 2003.
7. David L. Olson and Dursun Delen, "Data Mining Process" in Advanced Data Mining Technique, Springer, 2008.
8. Getoor, L. (2003). Link Mining: A New Data Mining Challenge. SIGKDD Explorations Volume 4, Issue 2.
9. Berry MJ, Linoff G. Data mining techniques: for marketing, sales and customer support. USA: Wiley, 1997.
10. **10.** Gupta, S., Kumar, D., & Sharma, A. (2011). Data Mining Classification Techniques Applied For Breast Cancer Diagnosis And Prognosis. Indian Journal of Computer Science and Engineering (IJCSE) 188-195.
11. Shouman, M., Turner, T., & Stocker, R. (2012). Applying k-Nearest Neighbour in Diagnosing Heart Disease Patients. 2012 International Conference on Knowledge Discovery (ICKD 2012) IPCSIT Vol. XX. Singapore:IACSIT Press
12. Berry, M.J.A. & Linoff, G.S. (1997). *Data Mining Techniques: For Marketing, Sales and Customer Support*. New York: John Wiley & Sons Inc.
13. Jantan, H., Hamdan, A. R., & Othman, Z. A., Human Talent Prediction in HRM using C4.5 Classification Algorithm, (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 08, 2010, 2526-2534.
14. Han, J. and M. Kamber, 2001. Data Mining: Concepts and Techniques. San Francisco, Morgan Kauffmann Publishers.
15. Miller, A., 1993. The application of neural networks to imaging and signal processing in astronomy and medicine. Ph.D. Thesis, Faculty of Science, Department of Physics, University of Southampton.
16. SE Brosette , AP Spragre, WT Jones and SA. Moser, "A data mining system for infection control surveillance", Methods Inf Med , Vol. 39, pp. 303-310, 2000.
17. Brewin, B. (2003). New health data net may help in fight against SARS. Computerworld , 37(17), 1, 59.
18. EliasLemuye, ―Hiv Status Predictive Modeling Using Data Mining Technology.
19. Milley, A. (2000). Healthcare and data mining. *Health Management Technology*, 21(8), 44-47.
20. Maja Hadzic, Fedja Hadzic and Tharam Dillon et al. Mining of patient data: towards better treatment strategies for depression. International Journal of Functional Informatics and Personalised Medicine, 2010
21. [Campbell2010] Kevin Campbell, N. Marcus Thygeson and Stuart Speedie. Exploration of
22. Classification Techniques as a Treatment Decision Support Tool for Patients with Uterine Fibroids*;* Proceedings of International Workshop on Data Mining for HealthCare Management, PAKDD-2010.
23. HianChyeKoh and Gerald Tan, ―Data Mining Applications in Healthcare‖, journal of Healthcare Information Management – Vol 19, No 2.
24. Berry MJ, Linoff G. Data mining techniques: for marketing, sales and customer support. USA: Wiley, 1997.
25. Brannigan, M. (1999). Quintiles seeks mother lode in health "data mining." *Wall Street Journal*, March 2, 1.
26. S. Viaene, A. Richard and D. G. Dedene, "A case study of applying boosting Naive Bayes to claim fraud diagnosis," IEEE Transactions on Knowledge and Data Engineering, vol. 16, no. 5, pp. 612–620, May 2004.
27. W. S. Yang and S. Y. Hwang, "A process-mining framework for the detection of healthcare fraud and abuse," Expert Systems with Applications, Article in Press, corrected proof, 2005.

28. David Page and Mark Craven, ―Biological Applications of Multi Relational Data Mining.

29. Silver, M. Sakata, T. Su, H.C. Herman, C. Dolins, S.B. & O'Shea, M.J. (2001). Case study: how to apply data mining techniques in a healthcare data warehouse. Journal of Healthcare Information Management, 15(2), 155-164.

30. Benko, A. & Wilson, B. (2003). Online decision support gives plans an edge. Managed Healthcare Executive, 13(5), 20.

31. Parameshvyas Laxminarayan, Sergio A. Alvarez, Carolina Ruiz, and Majaz Moonis, "Mining Statistically Significant Associations for Exploratory Analysis of Human Sleep Data", IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE, VOL. 10, NO. 3, JULY 2006

32. Amir R. Razavi & Hans Gill & Hans Ahlfeldt, Nosrat Shahsavar, "Predicting Metastasis in Breast Cancer: Comparing a Decision Tree with Domain Experts", J Med Syst (2007) 31:263–273

33. McQueen, G. & Thorley, S. (1999). Mining fool's gold. *Financial Analysts Journal*, 55(2), 61-72.

34. Hand, D.J. (1998). Data mining: statistics and more? *The American Statistician*, 52(2), 112-118.

35. Ceusters, W. (2001). Medical natural language understanding as a supporting technology for data mining in healthcare. In Medical Data Mining and Knowledge Discovery, Cios, K. J. (Ed.), Physica- Verlag Heidelberg, New York, 41-69.

36. Megalooikonomou, V. & Herskovits, E.H. (2001). Mining structurefunction associations in a brain image database. In Medical Data Mining and Knowledge Discovery, Cios, K. J. (Ed.), Physica-Verlag Heidelberg, New York, 153-180.

