International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 03 | Mar-2018 www.irjet.net p-ISSN: 2395-0072

Advances in Data Mining: Healthcare Applications

Rakhi Ray

Department of Computer Science and Engineering
Jessore University of Science and Technology (JUST), Jessore – 7408, Bangladesh

---

Abstract – *Owing to the great advantages various organizations are using data mining technology. Healthcare is a vital part for everyone. Different new technologies are inventing to examine physical conditions and finding symptoms of the different disease. There is a huge amount of data involved with it including a patient's past medical records, examination history, and even the personal details. On the other hand, in some cases, the symptoms are available before something happens to someone such as stroke. If the symptoms are known, anyone can take enough precaution, and the sudden risk of severe effect can be minimized or even can be avoided. Since there are large amount of data related to the medical systems, an efficient method to find the appropriate data from the database is required. Data mining is one the best solutions for this purpose. A study on data mining application in healthcare with some recent progress in this field is presented in this paper.*

Key Words: *Data mining, knowledge-discovery in database (KDD), healthcare, medical data, techniques, applications, data management.*

1. INTRODUCTION

A technique that mines data effectively from a huge database to get useful information is called data mining. Since data mining extracts valuable information from large data-set, it is useful for various applications, e.g., scientific areas and commercial [1]. Some of the commercial applications of data mining include banking sectors, such as detection of fraud and credit scoring, to schedule maintenances and quality control it used by various manufacturers. Furthermore, data mining is used extensively in markets for up-selling or cross-selling and direct marketing, and segmentation of market by retailer [2]. In the case of banking sectors, almost every bank has a central database, transactions are online, and updates real time and a huge amount of data is generating through the whole process. It is extensively hard or almost impossible to get important information manually for the decision makers from the huge database [3]. On the other hand, efficient data mining technology can ease the whole process. Data mining can bring significant benefit to language research and language engineering, distance learning and web-based education, software maintenance, sports data, the intelligence agencies, and digital library etc. [4].

Similar to the other sectors discussed above, in healthcare sector data mining plays a vital role. Important healthcare data can be extracted using data mining. Analyzing data from different hospitals can be beneficial to get an idea of the diseases present mostly, the main case for the disease, symptoms, precautions and their remedies. In this way, the various disease can be prevented, or their effect can be minimized [5]. Details of data mining application are given later.

In this paper, a survey on data mining for the application on healthcare sector is discussed. It is seen that several studies are available on data mining for healthcare application, in which most of them are published recently. After reviewing different studies, it is found that several medical related applications using data mining techniques such as hospital management, pharmaceutical industries, and medical device industries, etc. This paper aims to provide an overview of the data mining starting from its definition, to the application, how this technique works together with a summary some of the recent work in the healthcare sector.

2. OVERVIEW OF DATAMINING

As mentioned before, the extraction process of useful information and patterns from a huge database is called data mining. In this section, an overview of the data mining is discussed in brief.

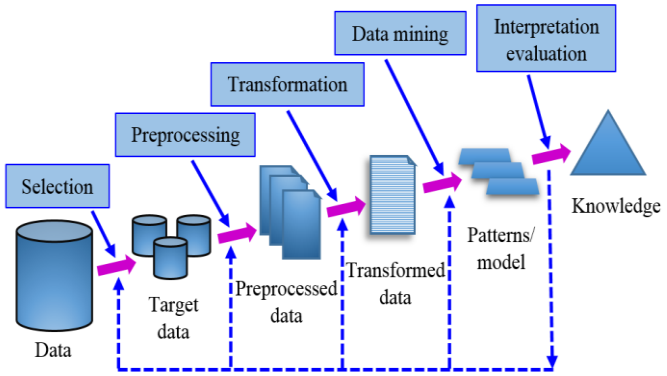

Fig -1: Knowledge discovery process.

2.1 Knowledge Discovery Process in Database

In the data mining process, computers can be trained to think like a human, through previous experience on a large database, to help human knowledge and reasoning skill-based decision making. It is also known as knowledge extraction from data, knowledge mining from data, knowledge discovery process or knowledge-discovery in database (KDD) [6]-[8]. Fig. 1 shows the knowledge discovery process from a database [3].

As shown in Fig. 1 the whole KDD process follows the steps discussed below [3], [4]

(i) Data Selection: In this step, a decision is made on the data which is relevant to the analysis, and those are retrieved from different locations.

(ii) Data Preprocessing: The data processing step itself consists of a combination of two different task. One of them is data cleaning, and the other one is data integration. Both are outlined below:

(a) Data Cleaning: In the data cleaning, also known as data cleansing, phase noise data and the data which is irrelevant to the analysis are removed from the collected data.

(b) Data Integration: In the data integration phase, multiple data sources are combined into a common source. It is worth mentioning here that the data sources are often heterogeneous.

(iii) Data Transformation: In the data transformation, also referred as data consolidation, phase the data which is already selected is transformed into forms to make it appropriate for mining procedure.

(iv) Data Mining: This step is the crucial step, clever techniques are applied in this phase, to extract potentially useful patterns.

(v) Interpretation and Evaluation: Based on given measures, interesting patterns representing knowledge are identified in this step.

(vi) Knowledge representation: This is the final phase of the KDD process. In this step, the discovered knowledge is presented visually to the user. Visualization techniques are used in this essential step to assist users to understand and interpret the results obtained from the data mining.

2.2 Data Mining Application Areas

New capabilities are required, which is not currently supplied by today's technology, for data mining as it is driven in part by new applications. These new applications can be divided into various categories [1]. As data mining is a relatively new technology, it is not matured fully yet. However, considering its benefit various organizations, mainly industries, are already using the data mining on a regular basis. Some of these include [6]

a. Banking sectors
b. Retail stores and e-commerce
c. Scientific, engineering and healthcare sectors
d. Insurance company
e. Sports organization

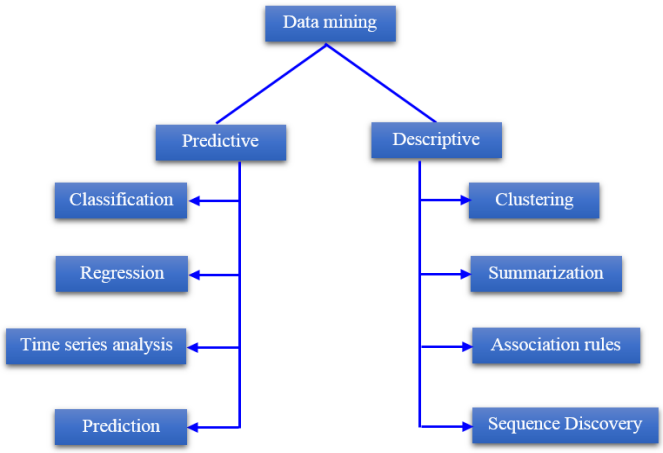

Fig -2: Data mining techniques.

2.3 Data Mining Techniques

Data mining tasks can be classified into two categories such as predictive and descriptive models as shown in the Fig. 2 [1]. Prediction of unknown data values by using known values are made in the predictive model. The descriptive model, on the other hand, is used to identify the relationships or patterns in the data and explores the properties of the examined data [9].

3. DATA MINING APPLICATION IN HEALTHCARE

As mentioned before, in healthcare sector there is a vast scope for data mining techniques to improve the medical science, and also the overall system. Nevertheless, research is medical science not only limited to the invention of new medicines (drugs) or advance instruments and techniques for disease identification, but also there are several other important things. For example, creating a data sheet for each patient including personal information, updating the entry on each visit etc. However, the creation of a patient profile, healthcare, diagnosis of disease is only a few examples of data mining application in healthcare [4]. Data mining in healthcare system indeed require significant effort because the data is complex, various types of data are related to healthcare system [10], [11]. Fuzzy based Neural Networks, Fuzzy logic, Genetic Algorithms, Artificial Neural Network, Nearest neighbor method, Decision trees, Bayesian Belief Networks, and Support Vector Machines are the commonly used techniques for data mining in healthcare sector [6],[11], [12].

The huge potential of data mining can be grouped in different ways. For example, healthcare management and inpatient length of stay prediction, effective treatment and diagnosis, detection of abuse and fraud, and relationship management. There are also some specialized data mining in medical technology including DNA micro-array analysis and medicine prediction. Some of the data mining applications mentioned above are discussed below in brief. Details are available in [2], [13] and the references therein.

3.1 Healthcare Management and Inpatient Length of Stay Prediction

Data mining can be applied to aid healthcare management system. It can be applied for better identification of high-risk diseases, for designing appropriate interventions, and to track chronic disease states. Furthermore using data mining technique, the number of hospital admission and claims can be reduced. Different medical centers are using data mining techniques to shorten patients' length of stay, minimizing clinical complications, improvement of medical practices, improvement of patients' outcomes and also providing information to the physicians. These eventually improve the quality of healthcare in a cost-effective manner.

3.2 Effective Treatment and Diagnosis

The effectiveness of medical treatments can be evaluated by developing data mining applications. To find the effectiveness of a course of action data mining can be applied. For example, by comparing courses of treatment provided patients, symptoms of the various disease, and contrasting causes data mining can find the effective treatment. As an example, to find most cost-effective yet best treatment can be decided by comparing the outcomes of a group of patients with the same condition or disease already treated using different drug regimens. It will also allow physicians to compare their practice patterns with others and also with peer-reviewed industry standards. In this way, data mining will enable successful standardization of specific disease treatment. In some cases, better diagnosis and treatment protocols are developed by comparing readmission and resource utilization data with recent scientific results from the literature.

3.3 Detection of Abuse and Fraud

Abuse of medical data and fraud can cost a lot of money and also related to some other issues, e.g., one's personal information. Applying data mining techniques, these can be identified. For example, using data mining technique abnormal or unusual patterns of claims by laboratories, clinics, physicians or from any other source can be identified. Inappropriate prescriptions, referrals, medical claims from insurance ban be detected using the data mining techniques. It was found that using data mining techniques a large amount of savings has been possible for different organizations through identifying a great number of suspects. Details are available in [2].

3.4 Customer Relationship Management

Although customer relationship is very familiar and vital aspect for commercial originations, such as retails and banking sectors, it is also very important in the healthcare system. Some of the customer interaction in the healthcare system are call centers, reception, billing departments, offices of the physicians, and inpatient settings. Data mining can be used in healthcare to improve the level of satisfaction by determining the usage patterns, future and current needs, and the preference of an individual. Moreover, the technique can be used to predict the purchase strategies of healthcare customer and hence pharmaceutical companies can also be benefited. In addition, this can bring ultimate satisfaction for each individual and also improve hospital profitability.

4. LITERATURE REVIEW

In this section, a review of some research of data mining application in the healthcare sector is given in brief including some recent research in this area.

In 2018, Graham, et al. proposed a method to predict admission in hospital from the emergency department (ED) for improving patient flow and stop overcrowding using data mining [14]. The risk of admission from the ED is predicted by using administrative data, 120 600 records collected routinely from two major acute hospitals in Northern Ireland, to compare contrasting machine learning algorithms. Three algorithms were used to build the predictive models. These are a logistic regression, decision trees, and gradient boosted machines (GBM). It was found that the GBM performs better than the other two algorithms. The outcome of the study could be beneficial for reducing overcrowding in the ED by improving patient flow and hence patient satisfaction can be improved.

In 2017, Zwick et al. proposed the data analysis on traumatic brain injury [15]. A probabilistic graphical modeling technique, also known as reconstructability analysis (RA) was used for this purpose. The approach of the study was to develop a dynamic model for brain trauma and a new TBI classification system, useful clinically, by discovering the unexpected relationships within the data and contributing to continuing efforts on the Brain Trauma Evidence Based Consortium (BTEC). Two main findings of the study are: (1) discriminating the severity of a concussion, a confounding variable for the Digit Symbol Test is education, and (2) improved performance on the Reaction Time test can be predicted from previous head injury.

To detect the heart failure (HF), Saqlain et al. proposed a multinomial Naïve Bayes (NB) algorithm in 2016 [16]. The dataset used a total of 30 variables. This algorithm is compared with various classification algorithms such as Neural Network, Logistic Regression, Random Forest and Decision Tree, SVM, were used to compare the proposed algorithm. Different parameters such as Precision, Accuracy, Recall and Area Under the Curve (AUC) were to measure the performance of the proposed algorithm. The model proposed in [15] reported achievement of AUC of 92.4% and an accuracy of 86.7%.

Data mining is able to answer complicated queries for diagnosing heart disease. The outcome of data mining can be beneficial to healthcare practitioners for making intelligent clinical decisions, which would be better than the traditional decision support systems [17]. The treatment cost can be minimized by providing effective treatments on time. Chikshe et al., studied different techniques in 2016, proposed for data mining to compare the best method for prediction of heart disease [17]. Since other algorithms like the decision tree,

Naïve Bayes, classification algorithms have some limitations, the authors' used a technique consisting of a k-mean algorithm and an Artificial Neural Network (ANN). A combination of k-mean algorithm reduces the excitation time.

In healthcare a large amount of data is available, and data mining technique can be used to extract different hidden information for the public healthcare data [18]. In 2016 Sharma et al. used different machine learning tools to analyze the healthcare data systems [18]. For example, RapidMiner and WEKA were applied for the analysis. For every classification techniques, the percentage of accuracy is used as a standard to measure the performance. For a particular data set, the appropriate technique was determined considering the highest accuracy of the applied techniques. For the study, the public healthcare dataset taken from HMIS portal of MoHFW was trained. The study found that for the particular analysis the Decision Tree algorithm using RapidMiner data mining tool suits best for the particular dataset. It was concluded that once the best classification method is known, getting efficient results will be helpful.

Data mining techniques are utilized in [19] for diagnosis and prognosis of cancer. One of the main reasons for human death is cancer, and early detection can be helpful for curing the disease. One of leading cancer for women is breast cancer which has seen increased significantly in past years. Early diagnosis of this increase the survival significantly, 97% of survival for five or more years [19]. Majali et al. proposed a system in 2016 using Classification and Association approach in data mining for diagnosis and prognosis of cancer. FP algorithm was used in Association Rule Mining (ARM) to achieve the patterns frequently observed in benign and malignant patients. Decision Tree algorithm under classification also used to predict the possibility of cancer in context to age. Wisconsin data set was applied to FP growth algorithm, and a rule was obtained that indicates the general behavior and range of values for malignant and benign tumor. Within various data mining classifier Decision tree was found to be the best predictor on Wisconsin dataset, and 94% class-labels were predicted correctly using this algorithm.

Considering the growing demand of blood bank sector, it is important to exploit the stored data efficiently. Sharma and Gupta proposed in 2012 that blood bank sector can be benefited greatly through data mining, which has the potential to analyze the gathered data in their information systems [20]. The number of blood donors was classified and predicted based on their blood group and age. For the study J48 algorithm and WEKA tool was applied. In this study real-world data, collected from an EDP department of a blood bank centre, was utilized. The experimental results of the study found that through training and evaluation an accuracy rate of 89.9% was obtained in the classification of blood donors.

5. CHALLENGES OF DATA MINING IN HEALTHCARE

Though data mining offers great benefit in the healthcare system, it is not a straightforward task. Some of the challenges of data mining applications in healthcare systems are discussed below [2], [13].

One of the major limitations of data mining in the healthcare is the relevant raw data is heterogeneous and voluminous. Data from various sources are associated with it, such as administrations, consultation of a patient with a practitioner, results from laboratories, interpretation and review of doctors, etc. Because of different settings and systems, the accessibility of data can be limited to data mining, and the process becomes complex for data collection, retrieval, and data analysis. However, any data should not be ignored since all the data components may have a considerable impact on diagnosis and progressions of a patient. So, before data mining, the data need to be collected. One of the approaches is to successfully build a data warehouse, which can be a time-consuming project as well as costly. An alternative of this is a distributed network topology for more efficient data mining.

Problem with data is another issue such as inconsistent or non-standardized data, corrupted or missing data. For example, different formats may use to record pieces of data in different sources. Without standard clinical vocabulary, it is indeed extremely difficult for data mining in healthcare. Poor mathematical characterization and non-canonical form of such high volume, complex and heterogeneous data is another barrier for successful data mining. There are some other vital issues related to the medical data such as data ownership, ethical issue, social and legal issues, etc. Another problem is because of large data, the results from data mining may found different significant and interesting patterns which may be useless. Knowledge in the domain area together with a proper understanding of data mining techniques are another requirement for successful data mining application. Furthermore, extensive investment is required for developing data mining technology in terms of time, money and effort. Data entry should be systematic and stored properly for future use. The main requirement of data mining is thorough planning, technological preparation work, be aware of the effectiveness of the technology and use it, and collaborative and co-operative work of everyone involved in data mining.

6. CONCLUSION

Nowadays various organizations are using data mining technique to reduce cost and simultaneously service quality improvement. These include business, telecommunication, sports, banking sectors etc. On the other hand, application of this emerging technology has not utilized properly in the healthcare sector. After studying a number of recently published research and review papers, it was found that data mining may bring significant benefit to the healthcare sector. The benefit not only include prediction of medical condition using the previous history of a patient from the database but also hospital management systems such as emergency division. This will also enable to take precautions for high-risk diseases by studying the symptoms from a database.

Though data mining technique in the healthcare is indeed complex its benefit is boundless.

REFERENCES

[1] M. Durairaj and V. Ranjani, "Data Mining Applications In Healthcare Sector: A Study," International Journal of Scientific & Technology Research, vol. 2, no. 10, pp. 29-53, 2013.

[2] H. C. Koh and G. Tan, "Data Mining Applications in Healthcare," Journal of Healthcare Information Management, vol. 19, no. 2, pp. 64-72, 2005.

[3] V. Bhambri, "Application of Data Mining in Banking Sector," International Journal of Computer Science and Technology, vol. 2, no. 2, pp. 199-202, 2011.

[4] S. P. Deshpande and V. M. Thakare, "Data Mining System and Applications: A Review," International Journal of Distributed and Parallel systems, vol. 1, no. 1, pp. 32-44, 2010.

[5] R. Karthiyayini and J. Jayaprakash, "Association Technique on Prediction of Chronic Diseases Using Apriori Algorithm," International Journal of Innovative Research in Science, Engineering and Technology, vol. 4, no. 6, pp. 255-259, 2015.

[6] M. Bharati and M. Ramageri, "Data Mining Techniques and Applications," Indian Journal of Computer Science and Engineering, vol. 1, no. 4, pp. 301-305, 2010.

[7] S. O. Hussien, S. S. Elkhatem, N. Osman, and A. O. Ibrahim, "A Review of Data Mining Techniques for Diagnosing Hepatitis", Sudan Conference on Computer Science and Information Technology (SCCSIT), pp.1-6, 2017.

[8] D. Verma and R. Nashine, "Data Mining: Next Generation Challenges and Future Directions," International Journal of Modeling and Optimization, vol. 2, no. 5, pp. 603-608, 2012.

[9] S. L. Nalawade and R. V. Kulkarni, "Application of Data Mining in Health Care," International Journal of Science and Research (IJSR), vol. 5, no. 4, pp. 262-268, 2016.

[10] P. Chantamit-o-pas and M. Goyal, "Prediction of Stroke Using Deep Learning Model" D. Liu et al. (Eds.): ICONIP, Part V, LNCS 10638, pp. 774–781, 2017.

[11] D. Tomar and S. Agarwal, "A Survey on Data Mining Approaches for Healthcare," International Journal of Bio-Science and BioTechnology, vol. 5, no. 5, pp. 241-266, 2013.

[12] N. P. Waghulde, N. P. Patil, "Genetic Neural Approach for Heart Disease Prediction," International Journal of Advanced Computer Research, vol. 4, no. 3, pp. 778-784, 2014.

[13] M. Pradhan, "Data Mining and Health Care: Techniques of Application," ISOI Journal of Engineering and Computer science, vol. 1, no. 1, pp. 18-26, 2014.

[14] B. Graham, R. Bond, M. Quinn, and M. Mulvenna, "Using Data Mining to Predict Hospital Admissions From the Emergency Department," IEEE Access, vol. 6, pp. 10458-10469, 2018.

[15] M. Zwick, N. Carney, and R. Nettleton, "Mining data on traumatic brain injury with reconstructability analysis," in Proc. IEEE Symposium Series on Computational Intelligence (SSCI), Honolulu, HI, pp. 1-6, 2007.

[16] M. Saqlain, W. Hussain, N. A. Saqib, Nazar, and M.A. Kha, "Identification of Heart Failure by Using Unstructured Data of Cardiac Patients," in Proc. 45th International Conference on Parallel Processing Workshops, pp.426-431, 2016.

[17] N. Chikshe, T. Dixit, R.Gore, and P. Akade, "Hybrid Approach for Heart Disease Detection Using Clustering and ANN," International Journal on Recent and Innovation Trends in Computing and Communication, vol. 4, no. 1, pp.119-122, 2016.

[18] T. Sharma, A. Sharma, and V. Mansotra, "Performance Analysis of Data Mining Classification Techniques on Public Health Care Data," International Journal of Innovative Research in Computer and Communication Engineering, vol. 4, no. 6, pp. 11381-11386, 2016.

[19] J. Majali, R. Niranjan, V. Phatak, and O. Tadakhe "Data Mining Techniques for Diagnosis and Prognosis of Cancer," International Journal of Advanced Research in Computer and Communication Engineering, vol. 4, no. 3, pp. 613-616, 2015.

[20] A. Sharma and P. C. Gupta, "Predicting the Number of Blood Donors through their Age and Blood Group by using Data Mining Tool," International Journal of Communication and Computer Technologies, vol. 1, no. 6, pp. 6-10, 2012.

BIOGRAPHY

Rakhi Ray was born in Jessore, Bangladesh. She received the B. Sc. degree in Computer Science and Engineering from Jessore University of Science and Technology (JUST), Jessore – 7408, Bangladesh in 2013. She worked as a Software Support Engineer at the True Services (Pvt) Ltd., Bangladesh. Her research interests include data mining, machine learning, e-commerce, artificial neural network, cyber security, and cloud computing.