A Comparative Study on Usage of Data Mining Techniques in Healthcare Sector

Ahsan Humayun
Research Scholar
Department of Computer Science
National Textile University,
Faisalabad Pakistan

Adeel Waqar
Research Scholar
Faculty of Science,
Engineering and Technology
Swinburne University of Technology,
Victoria Australia

ABSTRACT
Recent advancement in the data mining technique has provided a platform to numerous applications in healthcare sector. It has become an active research area due to its large scale potential. In this survey, we have tried to present the application areas of data mining approaches specifically in the healthcare sector. The different approaches of the data mining and where they are used is also described in the paper. In this work, we have also tried to explain the different challenges faced by the data mining techniques in order to implement them. This survey also covers the pros and cons of the data mining techniques in healthcare. Data mining techniques have a tremendous future and it should be taken at its earliest because of the significant importance of the healthcare issues.

Keywords
Data Mining, Healthcare, Classification, Regression, Clustering, Association, Data Mining Applications, Data Mining Challenges

1. INTRODUCTION
Data mining is considered as one of the most challenging and top most research area in healthcare due to high importance of healthcare issues. The recent advancement in the data mining approaches has provided a platform to numerous applications in healthcare sector. It has become a active research area due to its large scale potential. In healthcare, data mining is playing a vial role in different fields like fraud detection in health insurance, cheaper medical treatments availability for the patients, Disease diagnosing and finding its procurement methods. It also accommodates the researchers in the field of healthcare in development of effective policies, and different systems to prevent different types of disease. The data required regarding such systems can be the details of the patients, hospitals, diseases and their treatments. Data mining is very much helpful for the analysis of different factors which are responsible for diseases spreading around, like working environment, living conditions, food quality, and availability of clean water, health services and many others [1].

2. RELATED WORK
Data mining was discovered in mid of 1990's and it becomes a very helpful and powerful tool which was used to extract previous unknown data and useful information from huge data set. Generally we can say that data mining and Knowledge discovery almost are the same terms that are alternatively used in databases. Data mining is very effective in providing information regarding healthcare issues, which is helpful for making medical decisions for doctors and staff; they can easily predict the disease and can propose a treatment for that disease. It is observed that many of the organizations are using the data mining techniques such as classification, clustering and association to increase their productivity regarding their healthcare matters [2].

JayanthiRanjan [3] in their paper had worked that how data mining is effective in extracting the unknown patterns, they also described that how data mining provides a strong decision making in pharmacy industry. K. Srinivas,[4] have described the most of the possible ways and areas where the data mining techniques are applicable. They have mainly discussed the decision trees, rule based techniques and naïve bayes technique for healthcare data. Shweta Kharya [5] has particularly an important contribution in the data mining approaches for the diagnosis of breast cancer. And how it can be cured with an effective treatment, she has proved this by using different data mining approaches.

3. DATA MINING APPROACHES
In this paper, a comparative study is done to analyze the different data mining approaches for the healthcare applications.

Classification approach works by dividing data sample into classes. Classification predicts the class for data samples. For example patients can be categorized as 'High risk' and 'Low risk' while considering their health status. There are two methods of classification. i.e.: Multilevel and Binary. Classification is considered as one of the mostly used approach in healthcare industry. It is used for different purposes, like estimation of the cost of the treatments, diagnosis of the cancer in the patients and for different skin diseases [6].

K-Nearest Neighbor (K-NN) is a classifier that is used for discovering unknown points using know points. The real time usage of K-NN is finding the relationship among the disease and the factors which are responsible for the disease. The major application area of K-NN is the patients suffering from heart disease. It is also used for diagnose of thyroid [7].

Decision Tree (DT) is another classifier with an idea that they built up a tree like structure. They are specifically used for the analysis of the probabilities under different conditions. The major application area of decision tree in the healthcare industry is to find out the survival of a patient suffering from cancer [8].

Support Vector Machine (SVM) is also a classification technique and specifically used for the patients suffering from the diabetics. The basic theme behind the working is that it built up the two dimensional plane and measure the diabetic ratio of the patient [9].

Neural Network (NN) have a wide application area in the healthcare industry, it helps out in identifying the different disease like chest disease, lungs cancer, asthma and other. It

not only helps in diagnosing of the disease but also helps for the treatment of the disease.

Bayesian Method is based on a theory which is also called as Bayesian Classification, it is also under the Classification tab, and a major application area in the healthcare industry. Bayesian method helps to identify the factors for a particular disease. It helps to figure out the ratio of different factors that how much these factors are responsible for spreading that particular disease [10].

Regression is another data mining technique that is widely used in the health care organizations. It is basically used to find out the different functions and the variables of that functions and their relationship. The basic usage of the regression technique is used to identify the survival of the patient that is suffering from a particular disease. It can predict that how much time a person have that is suffering from cancer on the basis of the symptoms of the disease, and the health condition of the patient [7].

Clustering is an unsupervised approach of data mining. The basic idea behind the clustering is that it identifies the similar points between the data samples; it works opposite to the classification. The major area of clustering is use of genetic algorithm in the healthcare data. The major application area of clustering is that they are widely used in the hospitals for their resource management. They can manage the data of the hospital patients, like how long a patient stays in the hospital [11].

Association is another data mining technique that is widely used. It has a great contribution the health care field. It is extensively used in identifying the disease and the factors and relationship between them. It is extensively used by different healthcare organizations for various purposes [7].

4. APPLICATION AREA

Healthcare is always an important and critical concern as it is a matter of human livings. It has always a great importance in each aspect of field. There is a huge application area of Data mining approaches in healthcare sector. In this section we have tried to summarize the all applications area of data mining techniques in specifically healthcare sector.

Cancer Detection is now a day's considered as the most increasing disease and it has a major concern in health care issues. There are lots of methods to cure cancer, but it is quite difficult to monitor the continuous condition or status of a patient suffering from cancer. Results of different studies depicts that the cells affected by the cancer produce nitric oxide which cause harm and make a tumor [1].

Diabetes is another one of the rapidly increasing disease in the world. A lot of work is done on controlling the diabetes of the patients. Different data mining techniques on massive scale have been developed to control the diabetic conditions of the patients. With the help of data mining techniques we can control the diabetes of the patients [8].

Asthma Detection Millions of people are now a day's suffering from asthma. It is also a fast growing disease due to increasing level of pollution in the environment. To overcome the asthma problem data mining technique have a great impact, Regression are used for this purpose on a large scale in many environmental maintenance organizations [12].

The most important issue in the healthcare sector is the human errors. Many of the accidents are caused due to the human faults. Sometimes the patient is not precisely checked by the doctor and hence the prescription of the doctor goes wrong which can lead death of the patient. In order to prevent such kind of accidents data mining techniques are used. According to a survey almost 98000 patients die every year due to such kind of errors [3]**.**

Strokes are very serious and recovery from heart strokes is very difficult and crucial. It needs proper attention and consideration of doctor about the patient. In this regard, data mining techniques are used on a large scale to avoid such kind of activities [13].

Hospital Ranking is a major application area of the data mining techniques which helps out the government to rank out the hospitals on the basis of evaluation that has been done by using a data mining approach. Ranking of the hospitals are done on the basis of different parameters like facilities provided by the hospital and the staff availability [14].

5. DATA MINING CHALLANGES

One of the major challenges is to gather the accurate and the relevant and significant data. It is quite difficult for the organizations to obtain quality data. Due to complex nature of healthcare industry, it often becomes quite difficult because the patients are not willing to provide their data, they consider it the as their privacy concern, which can later lead towards serious consequences. Healthcare finance systems are responsible for maintain the patients data. In such kind of systems there are about 300 questions which are quite hectic to answer for a patient. Another difficulty that is faced by the data mining techniques is the data sharing. Healthcare organizations are not willing to share their data. They consider it the violation of the privacy of the patients [15].

6. CONCLUSION AND FUTURE DIRECTIONS

In this survey, we have concluded that there is not even a single classifier that gives the best results for every data sample. So a classifier is only selected when it is properly tested and its performance is best between all the classifiers. There are different validation methods for the testing of the classifiers, in future we need to develop better validation techniques to test the different classifiers. It is also concluded that clustering is only used in the case of less availability of the data samples. But still the problem arises that what type of clustering we need to use. This is a major concern that is being faced by the different clustering scenarios. To avoid such kind of issues, randomly generated data points should be used to acquire results, but due to this, uncertainty arises that the results are accurate or not [6].

In the end, we are unable to conclude a single data mining technique as a best or accurate technique which gives perfect or accurate results in each of the problem or the issue specifically in the health care industry. Data mining provides many benefits in the healthcare industry to every instance. It provides benefits to the doctors, management, hospital staff and organizations as well. It provides the effective treatment and the diagnosing of the disease in accurate time. It provides the health services in cheaper cost and minimizes the fraud in the healthcare industry.

7. REFERENCES

[1] M. Durairaj and V. Ranjani, "Data mining applications in healthcare sector a study," *International Journal of Scientific and Technology Research,* vol. 2, pp. 29-35, 2013.

[2] M. Silver, T. Sakata, H.-C. Su, C. Herman, S. B. Dolins, and M. J. O Shea, "Case study: how to apply data mining techniques in a healthcare data warehouse," *Journal of healthcare information management,* vol. 15, pp. 155-164, 2001.

[3] J. Ranjan, "Applications of data mining techniques in pharmaceutical industry," *Journal of Theoretical and Applied Information Technology,* vol. 3, pp. 61-65, 2007.

[4] D. Shukla, S. B. Patel, A. K. Sen, and P. K. Yadav, "Analysis of Attribute Association Rule from Large Medical Datasets towards Heart Disease Prediction," 2013.

[5] S. Kharya, "Using data mining techniques for diagnosis and prognosis of cancer disease," *arXiv preprint arXiv:1205.1923,* 2012.

[6] H. Broekhuizen, C. G. Groothuis-Oudshoorn, J. A. Til, J. M. Hummel, and M. J. IJzerman, "A review and classification of approaches for dealing with uncertainty in multi-criteria decision analysis for healthcare decisions," *Pharmacoeconomics,* vol. 33, pp. 445-455, 2015.

[7] D. Tomar and S. Agarwal, "A survey on Data Mining approaches for Healthcare," *International Journal of Bio-Science and Bio-Technology,* vol. 5, pp. 241-266, 2013.

[8] I. Yoo, P. Alafaireet, M. Marinov, K. Pena-Hernandez, R. Gopidi, J.-F. Chang, *et al.*, "Data mining in healthcare and biomedicine: a survey of the literature," *Journal of medical systems,* vol. 36, pp. 2431-2448, 2012.

[9] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine," in *International Workshop on Ambient Assisted Living*, 2012, pp. 216-223.

[10] R. Bhuvaneswari and K. Kalaiselvi, "Naive Bayesian classification approach in healthcare applications," *International Journal of Computer Science and Telecommunications,* vol. 3, pp. 106-112, 2012.

[11] P. Delias, M. Doumpos, P. Manolitzas, E. Grigoroudis, and N. Matsatsinis, "Clustering healthcare processes with a robust approach," in *26th European Conference on Operational Research*, 2013.

[12] H. C. Koh and G. Tan, "Data mining applications in healthcare," *Journal of healthcare information management,* vol. 19, p. 65, 2011.

[13] V. Chaurasia and S. Pal, "Early prediction of heart diseases using data mining techniques," *Carib. j. SciTech,* vol. 1, pp. 208-217, 2013.

[14] D. He, S. C. Mathews, A. N. Kalloo, and S. Hutfless, "Mining high-dimensional administrative claims data to predict early hospital readmissions," *Journal of the American Medical Informatics Association,* vol. 21, pp. 272-279, 2014.

[15] A. Holzinger, M. Dehmer, and I. Jurisica, "Knowledge discovery and interactive data mining in bioinformatics-state-of-the-art, future challenges and research directions," *BMC bioinformatics,* vol. 15, p. 1, 2014.