Analysis of Application of Data Mining Techniques in Healthcare

Kamna Solanki
Assistant Professor,
M.D University,
Rohtak, India

Parul Berwal
Research Scholar,
M.D University,
Rohtak, India

Sandeep Dalal
Assistant Professor,
M.D University,
Rohtak, India

Sudhir
Research Scholar,
M.D University,
Rohtak, India

ABSTRACT
Data mining is a growing research area in various fields due to its limitless approaches to mine the data in target oriented manner. With advance research in health sector, there is plentitude of data available in healthcare sector. But the common research problem is; how to use the existing healthcare information in a fruitful targeted way. To solve this problem, Data Mining is the best available technique. This paper makes review and analysis of different data mining techniques such as Clustering, classification, and Regression etc. used in health care sector..

Keywords
Data Mining , Clustering, Classification, Regression, Healthcare.

1. INTRODUCTION
In early 1970's, storage of data and extraction of information was a cumbersome task. But in late 80's, due to advancement in the field of technology, storage and extraction of data became easy due to ongoing research in Data mining. Data Mining is the process of analyzing data from different perspectives and summarizing it into useful information-information that can cut costs etc. Technically, data mining is the process of finding correlations or patterns in hundreds of fields in large relational databases,

Data mining consist of 5 major elements:

- Extract, transform and load data onto data warehouse system.
- Store and manage the data in multidimensional database.
- Provide data access to analysts.
- Analyze the data by application software.
- Present the data in useful format.

Steps in a Data Mining Process are:

1. Data integration: first of all, data is collected from and integrated from all different sources.
2. Data selection: data is selected on behalf of some criteria.
3. Data preprocessing: data collected may contain errors, inconsistencies that need to be removed.
4. Data transformation: the data even after preprocessing may not be ready for mining so it needs to be transformed into a form appropriate for mining.
5. Knowledge discovery: meaningful patterns are extracted from large data. At last, meaningful patterns help in decision making.

Data mining is vastly being used in solving numerous research problems. Data mining is becoming increasingly popular and essential in healthcare sector [1]. Data mining applications can provide advantage to all parties involved in the healthcare industry [2] [3]. For example, data mining can help healthcare insurer detect fraud and abuse, physicians identify effective treatments and best practices and patients receive better and more affordable healthcare services [4]. The huge amount of data generated by healthcare transactions is too complex and voluminous to be processed and analyze by traditional methods. Data mining provides the methodology and technology to transform these amounts of data into useful information for decision making in healthcare sector.

2. DATA MINING TECHNIQUES
2.1 Classification
Classification is the process of predicting output based on some given input data. The goal of classification is to accurately predict the target class for each case in the data [5].In order to predict the data, it processes the training set and predictive set. It first develop relationships between the attributes of training data set .Then it is provided with the predictive data set, which contains similar attributes but with different data values, Then it analyze the given data and produce prediction by placing the different data sets in different classes based on the relationship of attributes [6][7].

For example, in a medical database; the training set would have relevant patient information based on its previous records, whereas the prediction attribute is whether the patient has chances of heart attack as shown in Table 1 and Table 2.

Table1: Training Set

AGE	HEART RATE	BLOOD PRESSURE	HEART PROBLEM
62	79	145/70	Yes
35	82	115/75	Yes
79	65	110/68	No

Table 2: Predictive Set

AGE	HEART RATE	BLOOD PRESSURE	HEART PROBLEM
45	96	143/69	?
63	54	108/73	?
83	95	115/68	?

Classification uses predictive rules expressed in the form of IF-THEN rules where the first part (IF part) consist of conjunction of conditions and the second part(THEN part) predict a certain prediction attribute value that satisfy the first part.

Using the above example, a rule predicting the first row in the training set may be represented as follows:

IF (age=62 and heart rate>72) or (age>60 and blood pressure>140/70) then Heart problem=yes.

This technique provide 80% prediction rate, but the optimal solution is a rule with 100% prediction rate; which is very hard to achieve.

Following are the classification techniques used in health care:

2.1.1 Decision trees

Decision tree is similar to flow chart in which every non-leaf node denote a test on a particular attribute and every branch represent a outcome of the test. Root node is the topmost node in the decision tree. For example, with the help of readmission tree, we can decide whether a patient needs to be readmitted or not. Using Decision Tree, a decision maker can choose best alternative and traversal from root to leaf indicates unique class separation based on maximum information gain [8] [9]. Decision tree are self explanatory and easy to follow. Set of rules can also be constructed with the help of decision tree. Decision Tree can be considered as nonparametric method because there is no need to make assumptions regarding distribution of space and structure of classifier. Decision tree have several disadvantages. These are: Most of the algorithm like ID# and C4.5 require target attributes to have discrete values as decision tree use divide and conquer strategy. More the complex relationship among attributes lesser is the performance.

2.1.2 Support vector machines

Vladimir Vapnik first introduced idea of Support Vector Machine [10]. Its accuracy is better than all other available techniques. It was first introduced for binary classification problems; but it can be further extended to multi class problems. It creates hyper-planes to separate data points [11].

It can be implemented in 2 ways:

1. Mathematical programming
2. Using kernel functions

With the help of training data sets, non linear functions can be easily mapped to high dimensional space. This can only be possible using kernel functions like Gaussian, sigmoid etc.

2.1.3 Neural network

It was developed in 20th century. Neural network was regarded as the best classification algorithm before the introduction of decision tree and SVM which has far better results. This was the reason that encouraged NN as the most widely used classification algorithm in various bio-medicine and health care fields. For Example NN has been used as the algorithm supporting the diagnosis of diseases like cancer and predict outcomes. In NN, basic elements are nodes or neurons. These neurons are interconnected and within the network they work together to produce the output functions. They are fault tolerant as they are capable of producing new observations from the existing observations in those situations where some neurons within the network fail. An activation number is associated with each neuron and a weight is assigned to each edge with in the NN. The basic property of NN is that it can minimize the error by adjusting its weights and by making changes in its structure as it is adaptive in its nature. One major advantage of NN is that it can properly handle noisy data for training and can reasonably classify new type of data which is different from training data. There are also various disadvantages of NN. First, it require many parameters including the optimum no of hidden layer nodes that are empirically determined and its classification performance is very sensitive to parameters selected. Second, its training or learning process is very slow and expensive.

Table3 depicts the usage of Classification techniques in healthcare sector.

Table 3: Usage History of Classification Techniques in HealthCare Sector

Researcher	Technique used	Purpose
1. Hu et al. [12]	SVM, decision tree, bagging and boosting.	To analyze micro array data.
2. Huang et al. [13]	Hybrid SVM based diagnosis model	For breast cancer.
3. Khan et al. [14]	Decision tree	For breast cancer.
4. Chang et al. [15]	Integrated Decision tree	For skin diseases in adults

	model.	and children.
5. Curiac et al.[16]	Bayesian method	For psychiatric disease.
6. Moon et al. [17]	Decision tree algorithm	To characterize the smoking behavior among smokers by assessing their psychological health conditions and consumption of alcohol.
7. Chien et al. [18]	Hybrid decision tree classifier.	For chronic disease.
8. Shouman et al. [19]	K-NN classifier.	For heart disease.
9. Liu et al. [20]	Fuzzy-NN classifier.	For thyroid disease.
10. Er et al. [21]	Artificial Neural network	For chest disease.

2.1 Regression

Regression is a data mining technique that helps in identifying those functions that are useful in order to demonstrate the correlation among different variables. It is a mathematical tool and can be easily constructed using training data sets.

Regression can be classified into linear and non linear based on certain count of independent variables. In order to estimate association between two type of variable in which one is dependent variable and another one is independent variable, linear regression is used. One of the disadvantages of this technique is that it cannot be used for categorized data. The categorical data can be used with the help of logistic regression. Usage of Regression for health sector has been summarized in table 4.

Table 4: Usage History of Regression Techniques in HealthCare Sector

Researcher	Technique used	Purpose
1. Divya et al [22]	Weighted SV Regression	To provide better healthcare services by continuously monitoring patients.
2. Xie et al. [23]	Regression decision tree algorithm	To study number of hospitalization days.
3. Alapont et al. [24]	Linear regression	For effective utilization of hospital resources.

2.2 Clustering

It is an unsupervised learning technique which is different from classification technique (supervised learning method). It is best suited for large amount of data. It works by observing independent variables. The main task is to form clusters from large databases on the basis of similarity measure. Different types of clustering algorithms are defined in table 5 and various clustering algorithms used in health care are described in table 6.

Table 5: Types of Clustering Algorithms in HealthCare Sector

Technique	Description
1. Partitioned Clustering	With the help of 'n' data points maximum possible of 'k' clusters is obtained by relocating objects to 'k' clusters.
2. Hierarchical Clustering	Data points are partitioned in tree form either top-down or bottom-up.
3. Density-based Clustering	It can handle cluster of any arbitrary shape whereas above two can handle only spherical shape clusters.

Table 6: Usage History of Clustering Techniques in HealthCare Sector

Researcher	Technique used	Purpose
1. Chen et al. [25]	Hierarchical clustering	To analyze micro-array data.
2. Chipman et al. [26]	Hybrid Hierarchical clustering.	To analyze micro-array data.
3. Bertsimas et al. [27]	Clustering algorithm	To predict health care cost.
4. Peng Y et al. [28]	Clustering algorithm	To detect healthcare frauds.
5. Belciug et al. [29]	Hierarchical, partitioned and density based clustering.	Efficient utilization of healthcare resources.

3. ACCURACY ANALYSIS OF VARIOUS DATA MINING TECHNIQUES IN HEALTHCARE

There are various challenges in healthcare data that create serious obstacles in decision making:

a. Different healthcare organizations use different formats for storage of data.

b. Lack of standard form of data for storage.

c. Data sharing is another important challenge that creates problem. Both the medical organizations and the patients are not ready to share their private data.

d. Another is to build centralized data warehouse is very time consuming and costly process.

Table 7 represents the comparative accuracy analysis of various data mining techniques in healthcare sector. Figure 1 represents the corresponding comparative chart for accuracy level of various data mining techniques in healthcare sector. So, it is evident from table 7 and figure1 that Classification techniques are most widely used and have highest accuracy level among all other techniques in healthcare sector.

Table 7: Comparative Accuracy Analysis of Data Mining Techniques in HealthCare Sector

Sr. No.	Disease	Data Mining Technique	Algorithm	Accuracy level (%)
1.	Heart Disease	Classification	Naïve	60
2.	Cancer	Classification	Decision Table	97.77
3.	HIV AIDS	Classification, Association Rule mining	J48	81.8
4.	Blood Bank	Classification	J48	89.9
5.	Brain Cancer	Clustering		85
6.	Tuberculosis	Naïve Bayes Classifier	NN	78
7.	Diabetes	Classification	C4.5	82.6
8.	Kidney Dialysis	Classification	Decision making	75.97
9.	Dengue	Classification	C5.0	80
10.	Hepatitus C	Classification	Decision tree	73.2

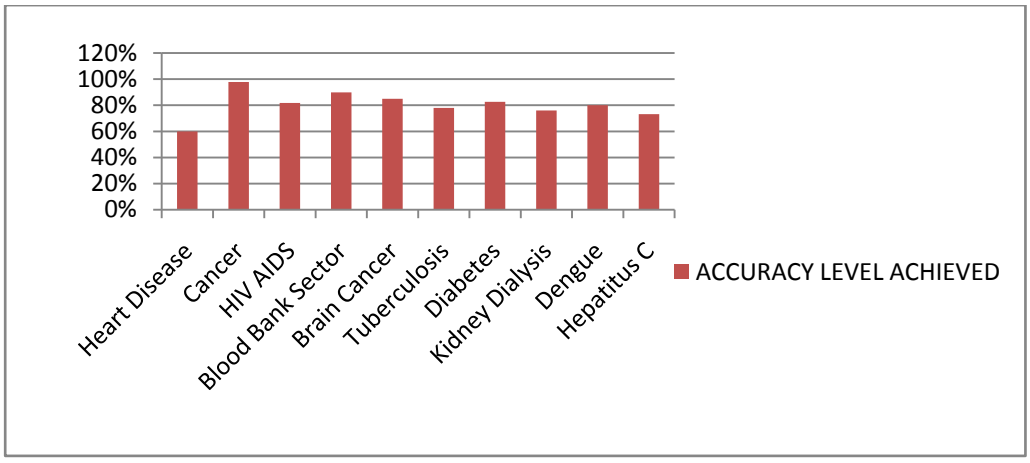

Figure 1 Chart for Accuracy Level Achieved in various disease in HealthCare Sector

4. CONCLUSION

As number of techniques of data mining are being used by various researchers in healthcare sector, so number of publications in this field are increasing. A tool is required to evaluate and summarize all the research work done so far in this particular field. A systematic literature review of all the data mining techniques used in health care is provided in this paper. Research identified a number of techniques used since 1970's. The paper summarizes all the techniques used in health care along with their accuracy level. A number of data mining techniques such as Decision tree classification, Support Vector Machine classification, Linear regression, Hierarchical clustering are the techniques that are mainly used by researchers as they provide high accuracy and efficiency.

The paper finally highlighted that the main goal of achieving high accuracy and efficiency which is very important in health care sector; still remains an open research issue. So any data mining technique that will solve this issue and will provide high accuracy and efficiency is the need of the hour. Our future research in this direction will try to propose a novel data mining technique that can provide better accuracy in wide variety of disease in comparison to peer available techniques.

5. REFERENCES

[1] T Koh, Hian Chye, and Gerald Tan, "Data mining applications in healthcare."*Journal of healthcare information management*, Vol. 19, No. 2, 2011, pp. 65-68.

[2] H. Kaur and Siri Krishan Wasan, "Empirical study on applications of data mining techniques in healthcare." *Journal of Computer Science* Vol. 2, No. 2, 2006, pp. 194-200.

[3] M. K. Obenshain, "Application of data mining techniques to healthcare data." *Infection Control & Hospital Epidemiology* Vol. 25, No.08, 2004, pp. 690-695.

[4] S. H. Liao, Pei-Hui Chu, and Pei-Yuan Hsiao, "Data mining techniques and applications–A decade review from 2000 to 2011." *Expert Systems with Applications*, Vol. 39, No.12, 2012, pp. 11303-11311.

[5] D. S. Deulkar and R. R. Deshmukh. "Data Mining Classification."*Imperial Journal of Interdisciplinary Research* Vol. 2, No.4, 2016.

[6] S. Palaniappan, and Rafiah Awang. "Intelligent heart disease prediction system using data mining techniques."IEEE Conference on *Computer Systems and Applications, 2008.*

[7] K. B. Srinivas, Kavihta Rani, and A. Govrdhan. "Applications of data mining techniques in healthcare and prediction of heart attacks." *International Journal on Computer Science and Engineering (IJCSE)* Vol. 2, No. 02, 2010, pp. 250-255.

[8] Parvez Ahmed, Saqib Qamar, and Syed Qasim Afser Rizvi. "Techniques of Data Mining In Healthcare: A Review." *International Journal of Computer Applications* Vol. 120, No.15, 2015.

[9] Durairaj, M., and V. Ranjani. "Data mining applications in healthcare sector a study." *International Journal of Scientific and Technology Research* Vol. 2, No.10, 2013.

[10] Vladimir Vapnik. "The support vector method of function estimation."*Nonlinear Modeling*. Springer US, 1998, pp.55-85.

[11] Christopher JC Burges, "A tutorial on support vector machines for pattern recognition." *Data mining and knowledge discovery* Vol. 2, No.2, 1998, pp. 121-167.

[12] Hu, Hong, et al. "A comparative study of classification methods for microarray data analysis." *Proceedings of the fifth Australasian conference on Data mining and analytics-Volume 61*. Australian Computer Society, Inc., 2006.

[13] Huang, Cheng-Lung, Hung-Chang Liao, and Mu-Chen Chen. "Prediction model building and feature selection with support vector machines in breast cancer diagnosis." *Expert Systems with Applications* Vol. 34, No.1, 2008, pp. 578-587.

[14] Khan, Muhammad Umer, et al. "Predicting breast cancer survivability using fuzzy decision trees for personalized healthcare." . *30th Annual IEEE International Conference in Engineering in Medicine and Biology Society, 2008*

[15] Chang, Chun-Lang, and Chih-Hao Chen. "Applying decision tree and neural network to increase quality of dermatologic diagnosis." *Expert Systems with Applications* , Vol. 36, No. 2 , 2009, pp. 4035-4041.

[16] Curiac, Daniel-Ioan, et al. "Bayesian network model for diagnosis of psychiatric diseases." *31st IEEE International Conference on* Information *Technology Interfaces, 2009.*

[17] Moon, Sung Seek, et al. "Decision tree models for characterizing smoking patterns of older adults." *Expert Systems with Applications* Vol. 39, No.1, 2012, pp. 445-451.

[18] Chien, Chieh, and Gregory J. Pottie. "A universal hybrid decision tree classifier design for human activity classification." *Annual IEEE International Conference of Engineering in Medicine and Biology Society (EMBC), 2012* .

[19] Shouman, Mai, Tim Turner, and Rob Stocker. "Applying k-nearest neighbour in diagnosing heart disease patients." *International Journal of Information and Education Technology* Vol.2, No.3, 2012, pp. 220.

[20] Liu, Da-You, et al. "Design of an enhanced fuzzy k-nearest neighbor classifier based computer aided diagnostic system for thyroid disease."*Journal of medical systems* Vol. 36, No.5, 2012, pp. 3243-3254.

[21] Er, Orhan, Nejat Yumusak, and Feyzullah Temurtas. "Chest diseases diagnosis using artificial neural networks." *Expert Systems with Applications, Vol.* 37, No.12, 2010, pp. 7648-7655.

[22] Divya Tomar, and Sonali Agarwal. "A survey on Data Mining approaches for Healthcare." *International Journal of Bio-Science and Bio-Technology* 5.5 (2013): 241-266.

[23] Xie, Yang, et al. "Predicting Days in Hospital Using Health Insurance Claims." *Biomedical and Health Informatics, IEEE Journal of* 19.4 (2015): 1224-1233.

[24] J. Alapont, et al. "Specialised tools for automating data mining for hospital management." *Proc. First East*

European Conference on Health Care Modelling and Computation. 2005.

[25] Chen, Tung-Shou, et al. "A combined K-means and hierarchical clustering method for improving the clustering efficiency of microarray." *International Symposium on Intelligent Signal Processing and Communication Systems, 2005*

[26] Chipman, Hugh, and Robert Tibshirani. "Hybrid hierarchical clustering with applications to microarray data." *Biostatistics* Vol. 7, No.2 2006, pp. 286-301.

[27] Bertsimas, Dimitris, et al. "Algorithmic prediction of health-care costs."*Operations Research* Vol. 56, No..6, 2008, pp. 1382-1392.

[28] Peng, Yi, et al. "Application of clustering methods to health insurance fraud detection." *IEEE International Conference on Service Systems and Service Management, 2006*.

[29] Belciug, Smaranda. "Patients length of stay grouping using the hierarchical clustering algorithm." *Annals of the University of Craiova-Mathematics and Computer Science Series* Vol. 36, No.2, 2009, pp. 79-84.

[30] Kavitha, K. S., K. V. Ramakrishnan, and Manoj Kumar Singh. "Modeling and design of evolutionary neural network for heart disease detection."*International Journal of Computer Science Issues* Vol.7, No.5, 2010, pp. 272-283.