Journal of Recent Research in Engineering and Technology, 3(9), Sep 2016, PP.01-11
ISSN (Online): 2349 –2252, ISSN (Print):2349 –2260

Research Article

A SURVEY ON DATA MINING APPROACHES FOR HEALTHCARE DOMAIN

M. Pavithra, Dr.R.M.S.Parvathi

Assistant.Professor in CSE Jansons Institute of Technology, Coimbatore, India.
Dean- PG Studies Sri Ramakrishna Institute of Technology, Coimbatore,India.

Received 20 June 2016; Accepted 23 July 2016

Abstract

In today's world, Healthcare is the most important factor affecting human life. Due to busy life ,it is not possible for personal healthcare. The proposed system acts as a preventive measure for determining whether a person is fit or unfit based on his/her historical and real time data by applying clustering algorithms viz. K-means and D-stream. Data Mining is one of the most motivating area of research that is become increasingly popular in health organization. Data Mining plays an important role for uncovering new trends in healthcare organization which in turn helpful for all the parties associated with this field. This survey explores the utility of various Data Mining techniques such as classification, clustering, association, regression in health domain. In this research, we present a brief survey of these techniques and their advantages and disadvantages. This survey also highlights applications, challenges and future issues of Data Mining in healthcare. Recommendation regarding the suitable choice of available Data Mining technique is also discussed in this paper.

Keywords:

Healthcare systems, Data Mining (DM), Clustering, classification, K-means clustering algorithm.

1.Introduction

Today the healthcare industry is one of the largest industries throughout the world. It includes thousands of hospitals, clinics and other types of facilities which provide primary, secondary & tertiary levels of healthcare. The delivery of health care services is the most visible part of any health care system, both to users and the general public. A health care provider is an institution or person that provides preventive, curative, promotional or rehabilitative health care services in a systematic way to individuals, families or community. In health care systems, the data mining is more popular and essential for all the healthcare applications. The

healthcare industry having more amounts of data, but this data have not been used properly for the application. In this health care system, data is converted into the useful purpose by using the data mining techniques .The data mining is the process of extracting or mining the knowledge from the large amounts of data, database or any other data base repositories.

Health Care Management is one of the most important and most popular research areas of the new millennium. The research studies on health care management aims both to control the increasing costs and to increase the accessibility level for health care services. Knowledge Management in Health care offers many challenges in creation, dissemination and preservation of health care knowledge using advanced technologies. Pragmatic use of Database systems, Data Warehousing and Knowledge

Management technologies can contribute a lot to decision support systems in health care. A formal definition of Knowledge discovery in databases is given as follows: "Knowledge Discovery in Data Mining is the non-trivial extraction of implicit previously unknown and potentially useful information about data".

Data mining technology provides a user-oriented approach to novel and hidden patterns in the data. The

Health care administrators to improve the quality of service can use the discovered knowledge. Clustering is considered as unsupervised classification technique, which aims at identifying the similar groups of entities available in given dataset. The entities in a subgroup (i.e. cluster) formed in clustering process are homogeneous to members of its own subgroup and heterogeneous to members of other subgroups (i.e. clusters). Although clustering has the ability to distinguish identical entities in a given dataset, but while dealing healthcare data, the process becomes more complex and difficult to obtain clusters, due to diversified and sparse nature of healthcare data. This work proposes an approach to cluster patients of similar clinical pathways using different clustering technique on a real healthcare dataset (data obtained provided by healthcare agency).

The reason behind the clustering techniques applied on healthcare data is that different groups of entities in certain medical treatment is highly complicated issue because each individual has its own medical history. Therefore, clustering algorithms working on different distance methods are exploited. For example, K-means algorithm works on partitioning method and agglomerative hierarchical clustering on hierarchical methods. The experiments are performed on real healthcare data for all considered clustering algorithms. The transactional healthcare data is firstly transformed into vectors representing physical diagnostic examinations of each patients, then different clustering techniques are

applied to achieve subgroups present in the data. The clustering results are finally evaluated by means of quality indexes.

2.Literature Survey-Data Mining in Healthcare Data

Data mining is widely used in several domains for identifying, detecting and analyzing data that is beneficial for future predictions, improving current processes and managing resources in optimal way. The benefits and limitations of IT in healthcare, and recommendations, as need of the hour to achieve desired outcomes, are reported. An integrated environment, which has provision of personalized healthcare services meeting the user specifications is also described .The safety of patients is one of the essential factors in HIT (Healthcare Information Technology) systems. Furthermore, HIT systems that lack in proper design and implementation may lead towards severe errors. Hence, enhanced approaches are recommended for HIT to deliver safer HIT services. The use of HIT system significantly improves service quality as well as reduces cost.

Clustering is greatly exploited in exploratory data analysis, pattern-analysis and grouping. A number of approaches are reported in literature for clustering healthcare data with respect to different aspects. For example, K-means algorithm is applied to cluster healthcare data, after transformation of binary data into real data. The transformation of data is made possible using Linear Wiener Transformation, which is normally used in noise filtering and is statistical transformation. The k-means cluster analysis has been incorporated into a methodology MCA (Multiple Correspondence Analysis), which investigates the characteristics of the people who use multiple healthcare resources. The proposed methodology helps in finding attributes of clustering in an optimal way. Furthermore, v-fold cross-validation is exploited in k-means cluster analysis to analyze the socioeconomic

and demographic characteristics of the people in the considered dataset, which are related to health care choices. The categorization of diabetic patients is carried out by hybrid model - three steps in cascaded fashion.

Firstly, incorrect classified instances are identified and removed by k-means clustering approach followed

by second step, where GA (Genetic Algorithm) and CFS (Correlation based Feature Selection) are applied for extracting relevant features. Finally, KNN (K-Nearest Neighbor) classifier is used for classification. The hybridized k-means clustering algorithm is proposed in which uses PCA (Principal Component Analysis) method on data and then k-means clustering is applied on resultant reduced data for analyzing high dimensional data. The experiments have been carried out on three different datasets of UCI machine learning repository: (i) Pima Indian Diabetes dataset, (ii) Breast Cancer dataset and (iii) SPECTF Heart dataset .A

fuzzy clustering technique that is based on symmetry has been developed to solve microarray data. A framework of subspace clustering has been proposed to examine the clustering of patient records for chronic diseases like diabetes and stroke.

3.Data Mining Challenges in Healthcare

One of the most significant challenges of the data mining in healthcare is to obtain the quality and relevant medical data. It is difficult to acquire the precise and complete healthcare data. Health data is complex and heterogeneous in nature because it is collected from various sources such as from the medical reports of laboratory, from the discussion with patient or from the review of physician. For healthcare provider, it is essential to maintain the quality of data because this data is useful to provide cost effective healthcare treatments to the patients. Health Care Financing Administration maintains the Minimum Data Set (MDS) which is recorded by all hospitals.

In MDS, there are many questions which are answered by the patients at check-in time. But this process is complex and patients face problem to respond the entire questions. Due to this MDS face some difficulties such as missing information and incorrect entries. Without quality data there is no useful results. For successful data mining, complication in medical data is one the significant hurdle for analyzing medical data. So, it is essential to maintain the quality and accuracy data for data mining

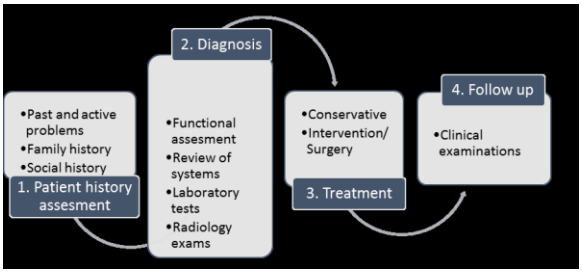

Fig 1 :Issues and Challenges of Data Mining in Medicine and Healthcare

The domain specific problems distinguish the process of DM in medicine and set it apart from other subject areas. R. Bellazi and B. Zupan stated that if DM were a simple process, the problems of information management would have been solved long ago. As the listed authors emphasized, the practical application of DM in medicine meets a number of barriers: technological, interdisciplinary communication, ethics, and protection of patient data. In addition, there are several well-known problems of biomedical data, such as inaccurate and fragmented information. Examples of inaccurate information measurements of vital functions were performed when the patient was not in a rest position; test sample required for testing was taken in non-sterile conditions; lab equipment calibration errors. Fragmented information includes cases, when the available patient data is non-sufficient for definitive results.

Another unique characteristic of DM in medicine is the usage in making decisions critical to human life. Therefore, as suggested by K. Cios et.al. the results of a selected DM method must be descriptive,

i.e. presented with explanations, so that medical experts can understand how these results were obtained. In terms of explicitness, some DM methods, such as decision trees, are more preferable than others, e.g. Neural networks.

Analysis of the data within the framework of several medical specialties raises additional challenges. In medicine, semantically the same concept may have multiple names and different identifiers in different code systems. Let us consider a hypothetical example. The department of anatomical pathology in a hospital uses Anatomical Pathology Laboratory Information Systems, in which SNOMED-CT nomenclature is used. While the cardiology department uses a cardiology information system, which has the ICD-10 and ICD-PT.

4. Clustering Techniques in Healthcare data

Clustering algorithms are widely adopted in several medical applications for different viewpoints. An approach for discovery and integration of frequent sets of features from distributed databases is presented in by means of unsupervised learning (i.e., hierarchical clustering). Precisely, frequent sets are extracted from distributed datasets and then are merged into a single frequent item set. Moreover, after applying hierarchical clustering, indexing is measured for quality results .Building analytical models of patients low in hospitals is demonstrated in using k-means clustering. emphasized on the data preparation phase as an essential step that affects the quality of solutions. Further, the size of the dataset is reported as main affecting factor to the solutions. The design of the care model for patients for a given collection is reported in .The approach in identities interested set of patients, then builds patients' care model (i.e., patterns of patients' care) and provides descriptions.

The criterion for similarity searching in medical databases does not differ from similarity searching in any other database. For example, in order to detect many diseases like Tumor, the scanned pictures or the x-rays are compared with the existing ones and the dissimilarities are recognized. Therefore, generalized clustering tools and techniques can be applied to medical databases, with little or no modification. The dataset used in the experiment contains records of blood tests of liver-disorder patients. The target is to cluster the patient’s records into different groups with respect to the test report attributes which may help the clinicians to diagnose the patient’s disease in efficient and effective manner. SOM toolbox (Matlab5) and k-means clustering algorithm have been used to group the database into different clusters.

The Self- Organizing-Map (SOM) is a powerful visualization tool based on vector quantization method which places the prototype vectors on a regular low dimensional grid in an ordered fashion [5]. The SOM Toolbox is an implementation of the Self-Organizing-

Map and its visualization in the Matlab5 computing environment. Since SOM Toolbox provides the visual view of the dataset and some obvious clusters can be immediately pointed out by a domain user, who can specify the input parameters for another powerful clustering algorithm named k-means to group the dataset in the found clusters. Figure 1 shows the scenario for a typical run of the clustering experiment. It represents the major activities, their sequential order and their interdependencies.

This also propose hierarchical K-means regulating divisive or agglomerative approach for better analyzing large micro-array data. It was reported that divisive hierarchical K-means was superior to hierarchical and K-means clustering on cluster quality as well as on computational speed. Apart from this, it was also mentioned that divisive hierarchical K-means establishes a better clustering algorithm satisfying researcher's demand .

5. Proposed Methodolgy

We proposed the hybrid hierarchical clustering approach for analyzing microarray data .In this research, the proposed hybrid clustering approach combines bottom-up as well as top-down hierarchical clustering concepts in order to effectively and efficiently utilizes the strength of both concepts for analyzing micro-array data. The proposed approach was built on a mutual cluster. A mutual cluster is a group of points closer to each other than to any other points. The research demonstrates the proposed technique on simulated as well as on real micro-array data. It used the approaches of classification trees and clustering algorithms in order to predict the cost of healthcare by using the dataset of three years collected from the insurance companies to perform the experiment. On the basis of analysis, following results were obtained in this research. First result shows that in order to provide accurate prediction of medical costs and to represent a powerful tool for prediction of healthcare costs data mining methods provide better accuracy. Another result shows that in order to predict the future costs pattern of past data was useful.

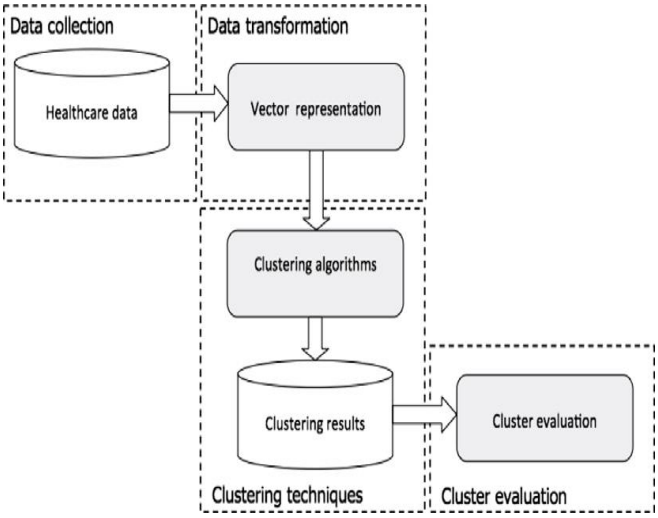

Fig.2. Proposed Approach of Clustering Healthcare Data.

It used the agglomerative hierarchical clustering approach for grouping the patients according to their length of stay in the hospital in order to provide better utilization of hospital resources and

provide better services to patients . Tapia et al., analyzed the gene expression data with the help of a new hierarchical clustering approach using genetic algorithm. In this research, the main focus was on regeneration of protein-protein functional interactions from genomic data. In this research, the proposed algorithm can predicate the functional associations accurately by considering genomic data. Soliman et al., proposed a hybrid approach for better analyzing the cancer diseases on the basis of informative genes.

The proposed approach used the K-means clustering with statistical analysis (ANOVA) for gene selection and SVM to classify the cancer diseases. On the basis of experiments that were performed on micro-array data, it has been found that the accuracy of K-means clustering with the combination of statistical analysis was better . Schulam et al., proposed a Probabilistic Subtyping Model (PSM) which was mainly designed in order to discovered subtypes of complex, systematic diseases using longitudinal clinical markers collected in electronic health record (EHR) databases and patient registries. Proposed model was a model for clustering time series of clinical markers obtained from routine visits in order to identify homogeneous patient subgroup. Belciug et al., concluded that among hierarchical, partitional, and density based clustering, the hierarchical clustering was provided effective utilization of hospital resources and provided improved patient care services in healthcare. Lu et al., proposed an Adaptive Benford Algorithm in the application area of healthcare insurance claims. The proposed algorithm was a digital analysis technique that utilizes an unsupervised learning approach in order to handle incomplete or missing data. This technique was applied to the detection of fraud and abuse in the health insurance claims using dataset, real health insurance data. The dataset was analyzed.

Table 1: Nomenclature of Diseases and Techniques used in Data Mining.

Disease	Technique Used
Conventional Pathology Data	Extracting patterns & detecting trends using Neural Networks [3].
Coronary heart disease	Prediction models using Decision Tree Algorithms such as ID3, C4.5, C5, and CART [3] [32].
Lymphoma Disease and Lung Cancer	Distinguish disease subtypes using Ensemble approach [4] [6].
Psychiatric Diseases	Predicate the probability of a psychiatric patient on the basis detected symptoms using BBN Bayesian networks [7].
Fre quent Disease	Identify frequency of diseases in particular geographical area using Apriori algorithm [8].
Liver diseases	Classification using Bayesian Ying Yang (BYY) [10].
Skin Disease	Categorization of skin disease using integrated decision tree model with neural network classification methods [14].
Diabetes	Classification of Medical Data using Genetic Algorithm [15].
Functional Magnetic Resonance Imaging (fMRI)	Integration of Clustering and Classification of biomedical databases [16].
Chest Disease	Constructed a model using Artificial Neural Network (ANN) [17].

6. Conclusion and Future Work

For any algorithm its accuracy and performance is of greater importance. But due to presence of some factors, any algorithm can greatly lost the above

mentioned property of accuracy and performance. Classification also belongs to such an algorithm. Classification algorithm is very sensitive to noisy data. If any noisy data is present then it causes very serious problems regarding to the processing power of classification. It not only slows down the task of classification algorithm but also degrades its performance. Hence, before applying classification algorithm it must be necessary to remove all those attributes from datasets which later on acts as noisy attributes. Feature selection methods play a very important role in order to select those attributes who improves the performance of classification algorithm.

Clustering techniques are very useful especially in pattern recognitions. But they suffer from a problem on choosing the appropriate algorithm because regarding datasets they do not have information. We can choose partitioned algorithm only when we know the number of clusters. Hierarchical clustering is used even when we do not know about the number of clusters. Hierarchical clustering provides better performance when there is less datasets but as soon as volume of datasets increases its performance degrades. To overcome this problem random sampling is very beneficial.

In hierarchical clustering, if the data is too large to be presented in a dendrogram, the visualization capability is very poor. One possible solution to this problem is to randomly sample the data so that users can properly understand the overall grouping/similarity of the data using the dendrogram that is generated with the sampled data. The main drawback to the use of hierarchical clustering algorithms is cubic time complexity. This complexity is such that the algorithms are very much limited for very large data sets. As the result, the hierarchical algorithms are much slower (in computational time) than partitioned clustering algorithms. They also use a huge amount of system memory to calculate distances between objects.

The privacy regarding to patient's confidential information is very important. Such type of privacy may be lost during sharing of data in distributed healthcare environment. Necessary steps must be taken in order to provide proper security so that their confidential information must not be accessed by any unauthorized organizations. But in situations like epidemic, planning better healthcare services for a very large population etc. some confidential data may be provided to the researchers and government organizations or any authorized organizations.

In order to achieve better accuracy in the prediction of diseases, improving survivability rate regarding serious death related problems etc. various data mining techniques must be used in combination.

To achieve medical data of higher quality all the necessary steps must be taken in order to build the better medical information systems which provides

accurate information regarding to patients medical history rather than the information regarding to their billing invoices. Because high quality healthcare data is useful for providing better medical services only to the patients but also to the healthcare organizations or any other organizations who are involved in healthcare industry. Takes all necessary steps in order to minimize the semantic gap in data sharing between distributed healthcare databases environment so that meaningful patterns can be obtained. These patterns can be very useful in order to improve the treatment effectiveness services, to better detection of fraud and abuse, improved customer relationship management across the world.

References

1.P.Santhi, V.Murali Bhaskaran "Performance of Clustering Algorithms in Healthcare Database", International Journal for Advances in Computer Science, Volume 2, Issue 1 March 2010

2.Vikram Singh, Sapna Nagpal "A Guided clustering Technique for Knowledge Discovery – A Case Study of Liver Disorder Dataset", International Journal of Computing and Business Research, Vol.1, no. 1, Dec 2010

3.Vijayarani S, Sudha S. "An efficient clustering algorithm for predicting diseases from hemogram blood test ", Journal of Science and Technology. 2015 Aug; 8(17):1–8.

4.Wasan Siri Krishan, Harleen Kaur, "Empirical Study on Application of Data Mining Techniques in Healthcare", Journal of Computer Science, 2(2), 194-200, (2010).

5.Suh S.C., Saffer S. and Adla N.K. "Extraction of Meaningful Rules in a Medical Database.",Proceedings of the 7th International Conference on Machine Learning and Applications, 450-456,2011.

6.Krishan W.S. and Kaur H. "Empirical Study on Application of Data Mining Techniques in Healthcare", Journal of Computer Science 2(2), 194-200,2010.

7.Wagstaff K., Cardie C., Rogers S. and Schrödl S,"Constrained k-means Clustering with Background Knowledge",Proceedings of the 18th International Conference on Machine Learning, 577-584, 2011.

8.Berks G., Keyserlingk D.G.V., Jantzen J., Dotoli M. and Axer H,"Fuzzy Clustering - A Versatile Mean to Explore Medical Databases. ESIT", Aachen, Germany, 453-457,2011.

9.Shouman, M.; Turner, T.; Stocker, R., "Using data mining techniques in heart disease diagnosis and treatment," Electronics,Communications and Computers (JEC-ECC), 2012 Japan-Egypt Conference on , vol., no., pp.173,177, 6-9 March 2012

10.Robu, R.; Hora, C., "Medical data mining with extended WEKA," Intelligent Engineering Systems (INES), 2012 IEEE

16th International Conference on , vol., no., pp.347,350,13-15June2012.

11.Khaleel et al.," A Survey of Data Mining Techniques on Medical Data for Finding Locally Frequent Diseases"International Journal of Advanced Research in Computer Science and Software Engineering 3(8), August - 2013, pp. 149-153

12.M. A. Masood, and M. Khan, "Clustering Techniques in Bioinformatics," International Journal of Modern Education and Computer Science (IJMECS), vol. 7, no. 1, pp. 38, 2015.

13.S.W. Purnami, J.M. Zain, & A. Embong, "Data Mining Technique for Medical Diagnosis Using a New Smooth Support Vector Machine," Communications in Computer and Information Science, 2010, pp.15-27,2011.

14.S. Kharya, "Using Data Mining Techniques for Diagnosis and Prognosis of Cancer Disease," International Journal of Computer Science, Engineering and Information Technology, vol.2, no.2, pp.55-66, 2012.

15.V.Bala Sundar, T.devi, N. saravanan "Development of a Data Clustering Algorithm for Predicting Heart", International Journal of Computer Applications (0975 – 888) Volume 48– No.7, June 2012.

16.Marina Gorunescu, Florin Gorunescu, Radu Badea, and Monica Lupsor "Evaluation on liver ¯broses stages using the k-means clustering algorithm", Annals of University of Craiova, Math. Comp. Sci. Series. Volume 36(2), Pages 19{24 ISSN: 1223-6934. 2009.

17.VelidePhani Kumar and Lakshmi Velide," Data Mining Approach for Prediction and Treatment Of diabetes Disease", IJSIT, 3(1), 073-079 , 2014.

18. Atul Kumar Pandey, Prabhat Pandey, K.L. Jaiswal, Ashish Kumar Sen ," DataMining Clustering Techniques in the Prediction of Heart Disease using Attribute Selection Method", International Journal of Science, Engineering and Technology Research (IJSETR) Volume 2, Issue 10, October 2013

19.G.Visalatchi, S.J Gnanasoundhari, Dr.M.Balamurugan," A Survey on Data Mining Methods and Techniques for Diabetes Mellitus", G.Visalatchi et al, International Journal of Computer Science and Mobile Applications, Vol.2 Issue. 2, pg. 100-105 , February- 2014.

20. Sathyabama Balasubramanian, Balaji Subramani," Symptom's Based Diseases Prediction In Medical System By Using K-Means Algorithm", International Journal of Advances in Computer Science and Technology Available Online at http://warse.org/pdfs/2014/ijacst13322014.pdf.,2010.

21.V. Manikantan & S. Latha , "Predicting the Analysis of Heart Disease Symptoms Using Medicinal Data Mining Methods", International Journal on Advanced Computer Theory and Engineering (IJACTE) ISSN (Print) : 2319 – 2526, Volume-2, Issue-2, 2013.

22.K. Rajalakshmi Dr. S. S. Dhenakaran," Analysis of Data mining Prediction Techniques in Healthcare Management System", Volume 5, Issue 4, April 2015 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering,2011.

23.Julia, A.M., and David, W.B., "Paperless Healthcare: Progress and Challenges of an it-enabled Healthcare System", Business Horizons,. Special Issue on Healthcare and the Life Sciences in Transition. Volume 53, No. 2, pp. 119-130, 2010.

24.Xuezhong, Z., Shibo, C., Baoyan, L., Runsun, Z., Yinghui, W., Ping, L., Yufeng, G., Hua, Z., Zhuye, G., and Xiufeng, Y., "Development of Traditional Chinese Medicine Clinical Data Warehouse for Medical Knowledge Discovery and Decision Support", Artificial Intelligence

in Medicine, Volume 48, No. 23, pp. 139-152, 2010.

25.Fahim, S., Ibrahim, K., and Abdun, N.M., "A Clustering Based System for Instant Detection of Cardiac Abnormalities from Compressed ECG", Expert Systems with Applications, Volume 38, No. 5, pp. 4705-4713, 2011.