A Survey on Data Mining Applications, Techniques and Challenges in Healthcare

[1]Dhruv Madan Gopal, [2]Aditya R, [3]C Vishnu Kumar Reddy, [4]Gautham S, [5]Nagarathna N
[1-4]UG Scholar, Dept. of C.S.E,
B.M.S College of Engineering, Bengaluru, India
[5]Associate Professor, Dept. of C.S.E,
B.M.S College of Engineering, Bengaluru, India

Abstract: An immense challenge facing the healthcare industry is the provision for quality services-correct, timely diagnosis and effective treatment, at affordable costs. Medical diagnosis is usually subjective. Also, the amount of data to be analysed for diagnostic purposes is huge and at times unmanageable. In this context, data mining can be used to efficiently infer patterns and rules from earlier treatments, thus helping to make diagnosis more objective and reliable. There is a huge amount of untapped data which can be analysed to obtain useful information through data mining. Here we have surveyed the various applications, technologies and challenges associated with data mining in healthcare.

Index Terms: Data Mining, Applications, Challenges, Diagnosis, Healthcare

I. INTRODUCTION

Data mining is the process of selecting, exploring and modelling large amounts of data. This process has become an increasingly pervasive activity in all areas of medical science research. It has resulted in the discovery of useful hidden patterns from massive databases. Data mining problems are often solved using different approaches from both computer sciences, such as multi-dimensional databases, machine learning, soft computing and data visualization; and statistics, including hypothesis testing, clustering, classification, and regression techniques. The use of data mining helps institutions make critical decisions faster and with a greater degree of confidence, and lowers the uncertainty in decision process. The integration of data mining can lead to the improved performance of Medical Decision Support Systems and can enable the tackling of new types of problems that have not been addressed before.

II. COMMON DATA MINING TECHNIQUES USED IN HEALTHCARE

The Data Mining techniques used in the healthcare sector are discussed in this topic.

[1] Discusses various techniques for early detection of heart diseases. They are- Support Vector Machine (SVM), Neural Networks, Naïve Bayes and Associative Classification.

Support vector machines are the supervised learning models with associated learning algorithms that analyse data and recognize patterns, used for classification and regression analysis. SVM constructs a hyper plane or set of hyper planes in high dimensional space which can be used for classification task.

Neural network is a set of connected input/output unit and weight associated with each connection. Network learns by adjusting weights so as to be able to predict the correct class labels of the input tuples. The neural networks are used in many applications like pattern recognition problems, character recognition, object recognition, and autonomous robot driving. There are many heart diseases prediction systems but none of them predict diseases base on risk factor like age, family history, obesity, alcohol intake etc.

Naïve Bayes is one of the successful classification methods. Classification is done based upon probability theory by computing prior Probability of the target attribute and conditional probability of remaining attribute. Naïve Bayes can't deal with continuous attributes so it converted into discrete by using equal frequency discretization method.

Associative classification is a new rule based approach that applies the methodology of association into classification. It adopt exhaustive search algorithm like Apriori and FP growth to generate the class association rule. It selects the small set of high quality rules from large number of rule to construct an efficient classifier. It is a two-step process first generate large no of rule using any associative rule mining algorithm then several rule pruning technique used to generate optimal rule set. There are two associative classification methods. Eager associative classification constructs a generalized model from training data set before receive any unknown instance for classification. After that new instance directly classify using the learned model. Lazy approach does not previously build a generalized model from training data but for each new instance to be classified, they process the stored training data samples. [2]Other techniques include rule induction and decision trees. Rule induction has a set of if-then-else rules, with two parts- a set of conditions (antecedents) and a set of associated results (consequents). It has the potential to use retrieved knowledge for prediction. Decision Tree is a method of knowledge representation organized like a tree with branches and nodes. Every node is

labelled with a class and branches coming out from an internal node have values of that node's attributes. This method is a common representation of instruction modelling. [3] We also frequently use Genetic Algorithms (GA) and Nearest Neighbour Method (NNM). GAs are modelled on the process of genetic modification, alteration and natural selection, inspired by the observation of evolution in nature. The algorithm creates a number of solutions for the given problem. Next, weaker solutions are discarded and the rest are preserved. Solutions can overlap. A good solution is hybridized and the process is again repeated till we have the best possible solution. These are used with association rules and other internal formulations to formulate hypotheses about dependencies between variables here. NNMs are used mainly for classification. There is no pattern used to categorize data, data given as input is the pattern. NNM chooses the subset of input data which is the best possible fit and forecasts accordingly. NNM is used to check accuracy of heart disease diagnosis. Studies shows that it achieved 97.5% accuracy.

In the next section, we will discuss the Applications of Data Mining that implement the Techniques discussed above.

III. APPLICATIONS OF DATA MINING IN HEALTHCARE

Here, we illustrate some of the Applications of Data Mining in Healthcare.

[4] has listed out some of the applications of data mining in the healthcare sector such as – insurance fraud and abuse detection, CRM decisions for an organisation, identifying effective treatment and best practices, providing patients with better and affordable services etc., She says that in spite of many approaches, the health care sector has more need for data mining today. There is a plethora of knowledge which can be gained from computerised health records but the vast amount of information stored makes it difficult. Application of data mining on medical data leads to discovery of new, useful and potentially lifesaving knowledge which would have otherwise been inert. Some current techniques are – telemedicine, Picture Archiving and Communication System (PACS), Digital Imaging and Communications in Medicine (DICOM), Electronic Medical Records (EMR) etc.

[5] Applications of Data Mining in the healthcare sector can be broadly classified into the following categories- Treatment Effectiveness, Healthcare management, Medical device industry, Pharmaceutical industry and system biology.

Now, we mention the Diagnostic Applications of Data Mining in the Healthcare Sector.

Diagnostic applications

Here we look at various technologies and techniques that are used for diagnostic purposes in the healthcare sector:

[6] Current expert diagnostic systems that employ data mining have a very low accuracy of prediction. This is because the techniques used are dependent on other records, they do not consider the patient's medical history and finally they are meant to be used only by domain experts and practitioners. To test this hypothesis, data model was built to predict the blood sugar level of diabetic patient, which also considered the medical history. This was tested on a dataset from UCI Machine Learning repository, Washington University, St. Louis, Missouri. There were around 10,000 records per patient, each having information such as age, glucose level etc. There was 11.26% improvement in the accuracy of prediction and 4.37% reduction in number of false cases.

[7] A framework which uses density based multiple level clustering i.e, performing multiple iterations over the data collection has been proposed. This deals with the inherent sparseness and variable distribution of data. In each iteration, a different part of the data is analysed and local clusters are identified for the set. The metric used is based on age, gender and examination history. By testing on a set of diabetic patients whose records by an Italian Health centre, the framework progressively identified clusters of patients with advanced stages. The first iteration yielded the patients with routine tests and subsequent iterations identified those with specific tests. The cluster had good silhouette values and prediction was 90% accurate.

[8] A data driven Model Predictive Control which finds a suitable duration of Hemo adsorption therapy for sepsis patients has been described. Here, therapy is applied in a non-continuous manner which saves 14% more patients than the usual method. According to this model, MPC is applied at each time point t to find the correct therapy for that time. If is recommended for that time, only one hour of treatment is administered. The patient is observed at the next point and so on. This was tested by looking through a population of non-survivor patients followed by a set of patients and temporal therapy to the patients. Finally, linear regression models for each variable are used on this training data. Non-continuous therapy cured around 41% with less than 12 hours of Hemo adsorption therapy and in some cases, two hours was enough.

[9] The use of a hybrid Rough–Genetic algorithm model which implements the advantages of Rough Set as an efficient and powerful analysis tool to identify the most relevant attributes has been discussed. Firstly, Rough Sets can be used to discover important and relevant facts hidden in datasets and express them with decision rules of natural language, and these results (rules) from a Rough-Set model are easily understood. Then, a Genetic Algorithm is used to optimize the rules induced using Rough Sets for classifying cases to test new medication for Hepatitis-C Virus (HCV) treatment. These algorithms encode a potential solution for a certain problem into a simple chromosome-like data structure, and then apply recombination operators to these data structures to preserve critical information. The experimental results obtained, show that the overall classification accuracy offered by the proposed Model is a dependable and superlative result.

[10] Predicting cardio vascular diseases based on Linear Discriminant Analysis of depression is possible. This also factors environmental variables like smoking, cholesterol levels, diabetes etc, for developing the prediction model. The dataset was obtained from the Korean National Health and Nutrition Examinations Survey (KNHANES) which had a total of 25,534 subjects, 335 depression patients were also included. Subjects were divided into training data and test data. Attributes considered are – sex, age, HDL cholesterol, total cholesterol, BP, smoking, diabetes and heart diseases. This information was applied to two equations for each patient. Equation 1 detects absence and equation 2 detects the presence of heart diseases. Classification is done depending on the higher value of the equations. Accuracy of FRS is 62.4% and for linear discriminant analysis, it is 69%.

[11] We can use the Group Method of Data Handling (GMDH) for predictive modelling of healthcare data. GMDH is a family of inductive algorithms for computer based mathematical modelling of multi parametric datasets. To find the best solution, GMDH looks at various models estimated by the method of least squares. It increases the number of partial model components and finds a model with optimal complexity. This is measured by an external criterion, the minimal value of which indicates optimal complexity.

[12] We can make use of Association Rule Summarizing techniques to detect the risk of diabetes mellitus. Common summarizing techniques are APRX – COLLECTION, RPGlobal, TopK and BUS. All techniques were applied to compress the original rule set of an electronic medical record to predict the risk of patients in the sub population. The most important differentiator between the techniques is the usage of a selection criterion to include a rule in the summary based on either expression of order or on the sub population covered by the rule. APRX and RPGlobal operate on expression while focussing on maximizing the compression, TopK and BUS operate on sub population with an aim to minimize redundancy. Association Rule Mining along with summarizing can help detect hidden clinical relationships and can also propose new patterns of conditions to redirect approaches for prevention, management and treatment. All four methods create reasonable summaries and each has its strengths.

[13] Co-clustering can be for diagnosis of heart diseases and also for detecting anomalies. Co-clustering acts as a powerful data analysis tool to diagnose heart disease heart disease and extract the data patterns of the datasets under test. Co-clustering, finds the subsets of rows in the dataset which are correlated with a subset of its columns. It is different from normal clustering that performs one way clustering such as k-means. On the other hand, in co-clustering simultaneous clustering of both row and columns happen. Co-clustering can produce a set of c column clusters of the original columns C and a set of r row clusters of original row instances R. Unlike other clustering algorithms, co-clustering also defines a clustering criterion and then optimizes it. In a nutshell, co-clustering finds out the subsets of rows and columns simultaneously of a data matrix using a specified criterion. From the summarization point of view, co-clustering provides significant benefits.

[14] Data mining techniques were used to show that Patient Characteristics are not associated with clinically important differential response to dapagliflozin. Baseline and early treatment response variable were selected and data mining methods have been used to rank all variables which are linked with reduction in glycated haemoglobin (HbAic) at week 26. Generalized linear modelling was then implemented using an independent set of data values to figure out which variables were predictive of dapagliflozin-specific treatment response as compared with the response of treatment in the control aim of the study. Finally, the simplest model was chosen by meta-analysis of nine other trials. This approach helped in minimizing risk of type1 errors. From a very big set of data, twenty two variables were used for generating the model as potentially predictive for dapagliflozin-specific reduction in HbA1c. Even though, baseline HbA1c was the variable that is most strongly associated with reduction in HbA1c at the end of the study, baseline fastring plasma glucose (FPG) was found to be the only predictive dapagliflozin-specific variable in the model. Placebo adjusted treatment effect of dapagliflozin and metformin vs metformin only for a change in HbA1c from baseline which was found to be -0.65% at the average baseline FPG of 192.3 mg/dL. This output turned down by 0.32% for every SD [57.2 mg/dL] rise in baseline FPG. The baseline FPG effect was confirmed in the meta-analysis of 9 studies, but its quality was smaller. But no other variable was predictive of dapagliflozin-specific reduction in HbAic independently. This study successfully identified a baseline reproducible predictor of differential response to dapagliflozin. Even when, the predictor was shown as baseline FPG, its magnitude was small to suggest clinical usefulness in identifying patients who benefit from the treatment of dapagliflozin treatment uniquely. Till date, this is one of the limited examples of methodologies that identify the predictive variable within conventional clinical datasets, as generating during last-stage clinical trials.

[15] There are many applications of relative neighbourhood graphs (RNG). RNG is frequently applied in machine learning, is used to decrease the size of training set. This reduction in size does not accompany a decrease in classification accuracy. It is applied to non-parametric classification rules in instance based learning. The classical *k*-nearest neighbour rule assigns an unknown instance by means of a majority vote amongst its *k* nearest neighbours, where *k* remains fixed after it is pre-tuned to the data at hand. The variation in RNG w.r.t the classical method is that it assigns an unknown instance based on majority vote amongst its neighbours. Here, the number of neighbours needed for a decision is not fixed in advance, and their choice does not depend only on the distances, but also on the local density and geometric structure of the data around the instance to be classified. RNG is used is a statistical test developed by Zighed, Lallich and Muhlenbach. They called the test separability index for application to supervised learning. This test was based on comparing relative weight of the edges in the RNG that connect data points of different classes, to the expected interval of a random distribution of the data labels on all the RNG edges. If the two values are not significantly different then no neighbourhood based method will yield a reliable prediction model.

[16] We can use an enhanced k-means clustering algorithm for discovery of patterns in medical data. The approach is to group a given information set through a certain number of groups (expect k groups) that have been established beforehand. The principle idea is to characterize k centroids, one for each group. These centroids have to be set in a guile manner resulting in a distinctive area of diverse effects. In this way, the better decision is to place them as far as possible from one another. The subsequent step is to take each point within a given information set and co-partner it with the closest centroid until reaching a state where all the points have been associated with a group. Once the first stage is done, and an unanticipated aggregating is carried out automatically, we need to reconfigure k new centroids as bary centers of each group due to the last step. After producing these k new centroids, another binding must be established between the same centroids set and the closest new centroid. A cycle will be produced. As an after effect of this cycle, we may recognize that the k centroids will change their regulated areas and at the end of the day centroids will not change their positions anymore. K means needs improvement with the initial random selection of the centroids array. The first step is to calculate all of the existing elements that have the highest degree in the space; from there we can have an initial configuration of what the clusters should look like. On the second run, we eliminate all the centroids that are in a single cluster and select *k* clusters with the highest results of the similarity function to be taken as the real cluster centroids. This being done, we iterate on the rest of the data elements to see if the centroids are going to change This will be performed exactly like the original *k*-means with both the distance and similarity functions.

[17] There is a cloud based healthcare application architecture titled eHealth cloud which uses a three tier architecture, each level having its own functionality. Tier-1 uses ria based client and enables the user to freely interact with the system. Secondly, the cloud server is simplified by using Amazon SimpleDB. Finally, the logic layer between client and server contains the rules for the system. There are three types of users – patient, doctor and administrator each having their own interface. It employs various data mining techniques for EMR.

These were the Diagnostic Applications. In the next section, we take a look at the other applications of Data Mining in the field of Healthcare.

There are also a number of other Data Mining Applications that will be listed out in the next section.

Other applications

In addition to diagnosis, Data Mining finds a variety of other applications in the healthcare sector.

[18] We have a decision support system using the predictive modelling which predicts and prevents strain situations in hospitals. There are 10 indicators of strain situations which are validated by professionals, one of which is Length of Stay (LOS). This makes use of several classification models and measures each one's performance using five metrics – Accuracy, Precision, Recall, Kappa Statistic and ROC. The dataset of 6,135 records between January and March 2012 was used. Exogenous values identified are – arrival time, age, tests etc. LOS is grouped into three types – < 289 minutes, 289 – 432 minutes, >432 minutes. Bayesian networks had the best precision (0.763), Kappa Statistic (0.36) and ROC (0.83). SVM had best accuracy (79.942%) and Recall (0.799).

[19] Similarly, there is a methodology using regression models to predict LOS of a new patient. Data is collected and formatted and the framework uses discrete event simulation to create new variables as per necessity. Identification of relevant variables is done using three approaches – hierarchical cluster analysis, attribute selection and principle component analysis. There are two linear models, both write the outcome as a sum of attribute values with appropriate weights assigned. The first model uses four values – Comp X-Ray, X-Ray, Echo and biology. The second model uses eight values – number of patients, AddressedBy , CAC, Echo, Scanner, X-Ray, AvisSpe, and Biology.

[20] Cross entropy can be used for detection of anomalous behaviour in health care services. Due to the very large quantity of information and the increased cost in health services, faulty behaviour may pass undetected and might be the reason of serious inefficiencies. Because the manual revision of data is costly and not convenient, other sources like data mining and anomaly detection has to be used to generate effective quantity and anti-corruption controls. There is only one way in determining if information is anomalous. That is to compare it with the information that is provided by other agents. This method can be misleading at times. To compensate this problem, we have to discriminate information that is given by each agent by a group of risk factors. The risk group can be basically told as a set of individuals that share the same diagnosis and also the identical socio-economic characteristics. The analysis is done separately in each of the risk groups. For every agent and every risk group, the cross-entropy of the agent's information is calculated. So, each and every risk group will have a measure of how faulty the information is, provided by each agent. Also, as the risk groups size increases information is also available in big volume. The Columbian health system is one of few existing system that use this mechanism. The results were found impressive and flexible with the fact that risk groups, variables can be defined with information from any year or any nation, and can be looked for anomalies.

Hence, these were the various Applications of Data Mining in Healthcare.

IV. CHALLENGES

In this topic, we mention the various challenges pertaining to the use of Data Mining in Healthcare.

[3] The main challenges of data mining in healthcare are – data generated is enormous, heterogeneous and from various sources which has an impact on diagnosis and should not be ignored. Secondly, interpretations of practitioners are in an unstructured language making it difficult to mine such data. There are challenges w.r.t. knowledge integrity assessment – proper expansion of efficient algorithms for different knowledge versions, algorithms for evaluating influence of modification of particular data on statistical importance of individual models. There is also restricted access to raw inputs. Future challenges are – improved data sharing between agencies, integrated web mining tools for text mining, standardisation and compression of data warehouses and representation and interpretation of findings.

[21] Integration of very large heterogeneous medical databases using data warehousing technologies has been discussed. Medical data is diverse in nature and found across many databases in various formats. This requires integration of data stored in such a way that it is consistent. Also, there is redundancy due to overlapping of information. Integration allows cross validation and verification of databases. There are totally four major components which are data sources, data staging area, data storage servers, data access. The data storage component incudes databases which act as sources of data supply to the warehouse. The data staging area is also a database which is intermediate in nature and is used for transferring data from source to warehouse. The data storage servers store the data in the warehouse. Data access component gives an interface for end users to retrieve data, to process, organize or analyse data and to export data to external sources. There are two major activities involved in implementing an MDW. The first is creating the structure of the warehouse with tables and the next is to design the tools end strategy for populating it. In the design process of the MDW, the first task is to determine the data structure to be used or what data will be included to the warehouse. Two solutions can be used which are need-based and availability-based approach. The need-based approach tests which data will be needed in the future based on the nature, so that these essential data can be collected and sent to the warehouse. On the other hand, the availability based approach examines which data is currently available in the operational systems. There will be some data which generally might not have an intermediate use but may find it useful in the future because of the fact that it is expensive to collect it later.
Thereby, the above challenges must be overcome to increase the efficiency of Data Mining in the field of Healthcare.

V. CONCLUSION

Data Mining is a very important and useful method which can be incorporated into the healthcare sector. Various techniques and algorithms are being proposed and employed for diagnosing different types of diseases. The use of data mining helps in discovering new patterns and knowledge, which may have previously been hidden and could potentially be lifesaving. Based on our findings, we aim to develop a system that efficiently and accurately diagnoses diseases based on symptoms obtained from the user.

VI. ACKNOWLEDGMENT

The work reported in this paper is supported by the college through the TECHNICAL EDUCATION QUALITY IMPROVEMENT PROGRAMME [TEQIP-II] of the MHRD, Government of India.

REFERENCES

[1]. Hadvani, K., & Limbad, N. (2014). A REVIEW ON EARLY HEART DISEASES PREDICTION USING DATA MINING TECHNIQUES.
[2]. Kaur, H., & Wasan, S. K. (2006). Empirical study on applications of data mining techniques in healthcare. *Journal of Computer Science*, *2*(2), 194.
[3]. Srinivas, B. S., Govardhan, A., & Kumar, C. S. (2014, January). Data Mining Issues and Challenges in Healthcare Domain. In *International Journal of Engineering Research and Technology* (Vol. 3, No. 1 (January2014)). ESRSA Publications.
[4]. Pradhan, M. (2014). Data Mining& Health Care: Techniques of Application.
[5]. Durairaj, M., & Ranjani, V. (2013). Data mining applications in healthcare sector a study. *Int. J. Sci. Technol.Res. IJSTR*, *2*(10).
[6]. Chandrakar, O., & Saini, J. R. (2014). Comparative Analysis of Prediction Accuracy of General and Personalized Datasets Based Classification Model for Medical Domain. *Heart Disease*, *1*, 1.
[7]. Xiao, X., & Chiusano, S. Analysis of Medical Treatments Using Data Mining Techniques.
[8]. Ghalwash, M., & Obradovic, Z. A Data-Driven Model for Optimizing Therapy Duration for Septic Patients. In *Proc. 14th SIAM Intl. Conf. Data Mining, 3rd Workshop on Data Mining for Medicine and Healthcare, Philadelphia, PA, USA (April 2014)*.
[9]. Eissa, M. M., Elmogy, M., Hashem, M., & Badria, F. A. (2014, April). Hybrid rough genetic algorithm model for making treatment decisions of hepatitis C. In*Engineering and Technology (ICET), 2014 International Conference on* (pp. 1-8). IEEE.
[10]. Yang, J., Lee, Y., & Kang, U. G. (2014). Cardiovascular disease prediction models on Linear Discriminant Analysis of depression.

[11]. SAHU, N. K., & SAHU, S. K. Predictive Modeling Technique in Data Mining for Health Care Data.
[12]. Kavitha, V., & Mohan, R. (2014). ARS: ASSOCIATION RULE SUMMARIZATION TECHNIQUES TO DETECT RISK OF DIABETES MELLITUS.
[13]. Ahmed, M., Mahmood, A. N., & Maher, M. J. Heart Disease Diagnosis Using Co-Clustering.
[14]. Bujac, S., Del Parigi, A., Sugg, J., Grandy, S., Liptrot, T., Karpefors, M., ... & Boothman, A. M. (2014). Patient Characteristics are not Associated with Clinically Important Differential Response to Dapagliflozin: a Staged Analysis of Phase 3 Data. *Diabetes Therapy*, *5*(2), 471-482.
[15]. Toussaint, G. T. (2014). Applications of the Relative Neighbourhood Graph.
[16]. Haraty, R. A., Dimishkieh, M., & Masud, M. (2015). An Enhanced k-Means Clustering Algorithm for Pattern Discovery in Healthcare Data. *International Journal of Distributed Sensor Networks*.
[17]. Biswas, S., Anisuzzaman, T. A., Kaiser, M. S., & Mamun, S. A. Cloud Based Healthcare Application Architecture and Electronic Medical Record Mining: An Integrated Approach to Improve Healthcare System.
[18]. Benbelkacem, S., Kadri, F., Chaabane, S., & Atmani, B. (2014, November). A DATA MINING-BASED APPROACH TO PREDICT STRAIN SITUATIONS IN HOSPITAL EMERGENCY DEPARTMENT SYSTEMS. In *10ème Conférence Francophone de Modélisation, Optimisation et Simulation-MOSIM'14*.
[19]. Combes, C., Kadri, F., & Chaabane, S. (2014, November). PREDICTING HOSPITAL LENGTH OF STAYUSING REGRESSION MODELS: APPLICATION TO EMERGENCY DEPARTMENT. In *10ème ConférenceFrancophone de Modélisation, Optimisation et Simulation-MOSIM'14*.
[20]. Villegas, A. R., Gómez, S. C., & De Arteaga, M. Cross-Entropy For Detecting Anomalous Behaviour In Health-Care Service.
[21]. Kumar, M. R. S. (2014). INTEGRATION OF VERY LARGE HETEROGENEOUS DATABASE FOR MEDICAL DATABASES BY USING DATAWARE HOUSING TECHNOLOGY.