SURVEY OF DATA MINING TECHNIQUES USED IN HEALTHCARE DOMAIN

Sheenal Patel and Hardik Patel

Department of Computer Science and Applications, Charotar University of Science & Technology, Changa, Gujarat, India

ABSTRACT

Health care industry produces enormous quantity of data that clutches complex information relating to patients and their medical conditions. Data mining is gaining popularity in different research arenas due to its infinite applications and methodologies to mine the information in correct manner. Data mining techniques have the capabilities to discover hidden patterns or relationships among the objects in the medical data. In last decade, there has been increase in usage of data mining techniques on medical data for determining useful trends or patterns that are used in analysis and decision making. Data mining has an infinite potential to utilize healthcare data more efficiently and effectually to predict different kind of disease. This paper features various Data Mining techniques such as classification, clustering, association and also highlights related work to analyse and predict human disease.

KEYWORDS

Data Mining, Health Care, Classification, Clustering, Association

1. INTRODUCTION

Data mining is an assortment of algorithmic techniques to extract instructive patterns from raw data. Healthcare industry today produces huge amounts of multifarious data about hospitals, resources, disease diagnosis, electronic patient records, etc. The large amount of data is crucial to be processed and scrutinized for knowledge extraction that empowers support for understanding the prevailing circumstances in healthcare industry. Data mining processes include framing a hypothesis, gathering data, performing pre-processing, estimating the model, and understanding the model and draw the conclusions [2]. Before studying how data mining algorithms are being applied on medical data, let us understand what types of algorithms exists in data mining and how they are functioning.

It came into existence somewhere in the middle of 1990's and appeared as a strong tool that extracts needful information from a bulk of data. In common, Knowledge Discovery (KDD) and Data Mining are related terms and are used interchangeably but several researchers assume that both terms are dissimilar as Data Mining is one of the most vital stages of the KDD process. According to Fayyad et al., the Knowledge Discovery in database is systematized in various stages whereas the first stage is selection of data in which data is gathered from different sources, the second stage is pre-processing the selected data, the third stage is transforming the data into suitable format so that it can be processed further, the fourth stage consist of Data Mining where suitable Data Mining technique is applied on the transformed data for extracting valuable information and evaluation is the last stage as shown in Figure 1 [28].

DOI : 10.5121/ijist.2016.6206

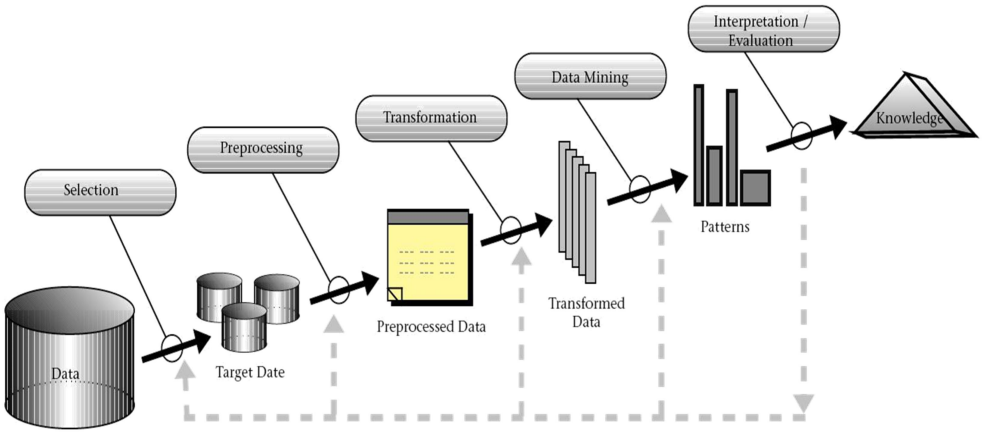

Figure 1. Stages of Knowledge Discovery Process

Knowledge Discovery in databases is the process of retrieving high-level knowledge from low-level data. It is an iterative process that comprises steps like Selection of Data, Pre-processing the selected data, Transformation of data into appropriate form, Data mining to extract necessary information and Interpretation/Evaluation of data [20].

Selection step collects the heterogeneous data from varied sources for processing. Real life medical data may be incomplete, complex, noisy, inconsistent, and/or irrelevant which requires a selection process that gathers the important data from which knowledge is to be extracted.

Pre-processing step performs basic operations of eliminating the noisy data, try to find the missing data or to develop a strategy for handling missing data, detect or remove outliers and resolve inconsistencies among the data.

Transformation step transforms the data into forms which is suitable for mining by performing task like aggregation, smoothing, normalization, generalization, and discretization. Data reduction task shrinks the data and represents the same data in less volume, but produces the similar analytical outcomes.

Data mining is a main component in KDD process. Data mining includes choosing the data mining algorithm(s) and using the algorithms to generate previously unknown and hypothetically beneficial information from the data stored in the database. This comprises deciding which models/algorithms and parameters may be suitable and matching a specific data mining method with the general standards of the KDD process. Data mining methods includes classification, summarization, clustering, regression, etc. [20]

Interpretation/ Evaluation step includes presentation of mined patterns in understandable form. Various types of information need different type of representation, in this step the mined patterns are interpreted. Evaluation of the outcomes is prepared with statistical justification and significance testing.

Knowledge discovery: integrating the extracted knowledge into another system for further action, or merely documenting the same and broadcasting it to interested parties. This step also comprises checking and resolving possible conflicts with previously extracted knowledge. [29] KDD can be effective at working with bulky data to define significant pattern and to develop strategic results. A health care organization can implement Knowledge Discovery in databases (KDD) by the help of experienced employee who has good understanding in health care domain [5].

Generally data mining algorithms are classified in two categories: descriptive model (or unsupervised learning) and predictive model (or supervised learning). Descriptive data-mining model is to discover patterns in the data and identifies the associations between attributes represented by the data. In contrast, the purpose of Predictive mining model is largely to predict the future outcome than existing behaviour [19].

2. DATA MINING TECHNIQUES

Data mining techniques such as association, classification and clustering are used by healthcare organization to increase their capability for building appropriate conclusions regarding patient health from raw facts and figures [24].

2.1. Classification

Classification comprises of two footsteps: - 1) Training and 2) Testing. Training builds a classification model on the basis of training data collected for generating classification rules. The IF-THEN prediction rule is highly popular in data mining; they signify facts at a high level of abstraction. The accuracy of classification model hinge on the degree to which classifying rules are true which is estimated by test data [9]. In health care domain classification can be made useful as "if DiabeticFamilyHistory=yes AND HighSugerIntake=yes THEN DiabetesPossiblity=High". Hatice et al., to analyse skin diseases by using weighted KNN classifier [1].

2.2. Clustering

Clustering is different from classification; it does not have predefined classes. A large database is divided into number of small subgroups called clusters. It divides the data based on similarities it have. Clustering algorithms discovers collections of the data such that objects in the same cluster are more identical to each other than other groups [13]. Tapia et al. examined the gene expression data with support of hierarchical clustering approach by using genetic algorithm [11].

2.3. Association

Association also has great impact in the health care industry to discover the relationships between diseases, state of human health and the symptoms of disease. Ji et al., used association in order to learn uncommon casual relationships in Electronic health databases [12]. An integrated approach of using Association and Classification techniques also improved the capabilities of Data Mining. Soni et al., have used this integrated approach of association and classification for studying health care data. This integrated approach is useful for determining rules in the database and then by using these rules, an effective classifier is raised. The study made experiment on the data of heart patients and generate rules by weighted associative classifier [26]. Thus, Association also has an ample influence in the healthcare field to identify the relationships among various diseases, state of human health and the symptoms of disease.

3. APPLICATION OF DATA MINING TECHNIQUES IN HEALTH CARE

The different classification algorithms mentioned below in figure 1 are used to predict or to analyse various diseases.

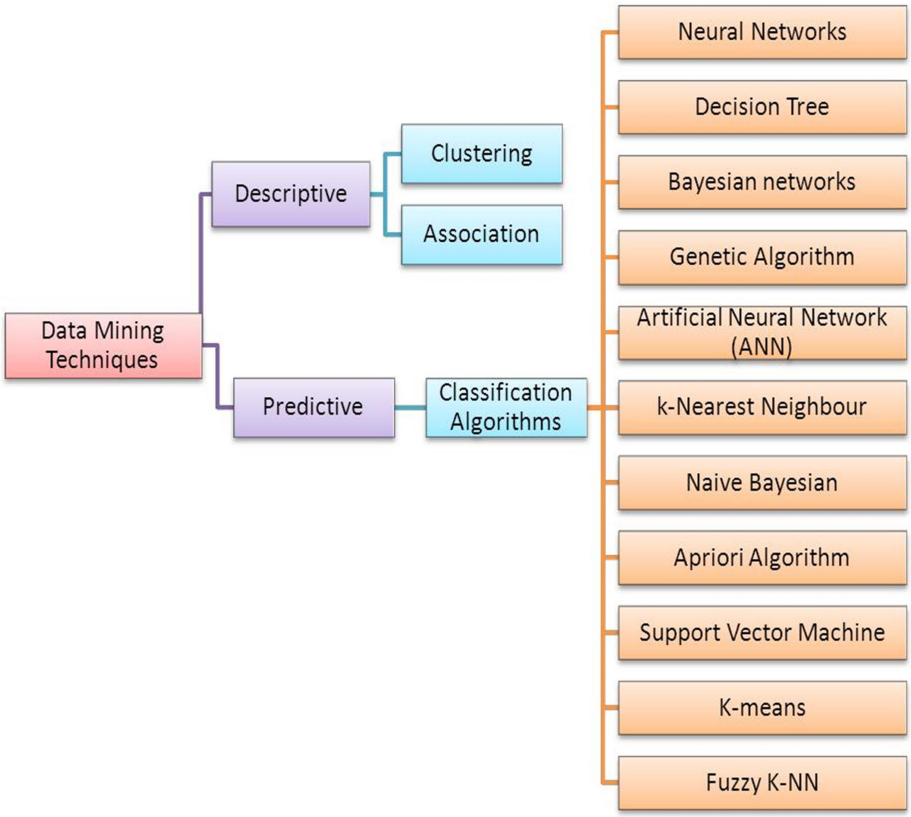

Figure 2. Different techniques in Healthcare domain

Summary of Techniques for Medical data mining.

In terms of prediction and decision making, Data mining techniques have substantial expansion in medical industry with respect to various diseases like diabetes, heart disease, liver diseases, cancer and others. Table 1 summarizes the medical data mining, its techniques used and for the related disease.

Table 1. Summary of medical data mining techniques

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].
Diabetes, Cancer	Classification of Disease using k-Nearest Neighbour [18].
Coronary Heart Disease	Improving classification accuracy using Naive Bayesian [21] [30].
Chronic Disease	Prediction of Diseases Using Apriori Algorithm [22].
Diabetes	Disease classification using Support Vector Machine [23]
Breast Cancer	Accurate Classification of medical data using K-means, Self-Organizing Map (SOM) and Naïve Bayes [25].
Cardio Vascular Diseases	Diagnose Cardio Vascular Disease using Classification algorithm [27].
Parkinson Disease	Familiarized an adaptive Fuzzy K-NN approach for diagnosing the disease [31]

4. CONCLUSION

With the recent rapid rise in the quantity of biomedical data that is gathered by electronic means in critical care and the rampant availability of inexpensive and dependable computing equipment, many researchers has started, or are eager to start, exploring these data. In this paper we observe some data mining techniques that has been employed for medical data. As there is voluminous records in this industry and because of this, it has become requisite to use data mining techniques to help in decision support and prediction in the field of Healthcare to identify the kind of disease. The medical data mining produces business intelligence which is useful for diagnosing of the disease. This paper throws light into data mining techniques that is used for medical data for various diseases which are identified and diagnosed for human health.

REFERENCES

[1] C. Hattice & K. Metin, "A DIAGNOSTIC SOFTWARE TOOL FOR SKIN DISEASES WITH BASIC AND WEIGHTED K-NN", Innovations in Intelligent Systems and Applications (INISTA), 2012.

[2] Dhanya P Varghese & Tintu P B, "A SURVEY ON HEALTH DATA USING DATA MINING TECHNIQUES", International Research Journal of Engineering and Technology (IRJET), Volume: 02 Issue: 07, Oct-2015.

[3] Doron Shalvi & Nicholas DeClaris, "AN UNSUPERVISED NEURAL NETWORK APPROACH TO MEDICAL DATA MINING TECHNIQUES", IEEE, 1998.

[4] Gustavo Santos-Garcia & Gonzalo Varela & Nuria Novoa & Marcelo F. Jimenez, "PREDICTION OF POSTOPERATIVE MORBIDITY AFTER LUNG RESECTION USING AN ARTIFICIAL NEURAL NETWORK ENSEMBLE", Artificial Intelligence in Medicine 30:61–69, 2004.

[5] Harleen Kaur & Siri Krishan Wasan, "EMPIRICAL STUDY ON APPLICATIONS OF DATA MINING TECHNIQUES IN HEALTHCARE", Journal of Computer Science 2 (2): 194-200, 2006.

[6] Hojin Moon & Hongshik Ahn & Ralph Kodell & Songjoon Baek & Chien- Ju Lin & James Chen, "ENSEMBLE METHODS FOR CLASSIFICATION OF PATIENTS FOR PERSONALIZED MEDICINE WITH HIGH-DIMENSIONAL DATA". Artificial Intelligence in Medicine 41:197–207, 2007.

[7] I. Curiac & G. Vasile & O. Banias & C. Volosencu & A. Albu, "BAYESIAN NETWORK MODEL FOR DIAGNOSIS OF PSYCHIATRIC DISEASES", Proceedings of the ITI 2009 31st Int. Conf. on Information Technology Interfaces, Cavtat, Croatia, 22-25 June-2009.

[8] Ilayaraja & T. Meyyappan, "MINING MEDICAL DATA TO IDENTIFY FREQUENT DISEASES USING APRIORI ALGORITHM", Proceedings of the 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering, 21-22 February-2013.

[9] Illhoi Yoo & Patricia Alafaireet & Miroslav Marinov & Keila Pena-Hernandez & Rajitha Gopidi & Jia-Fu Chang & Lei Hua, "DATA MINING IN HEALTHCARE AND BIOMEDICINE: A SURVEY OF THE LITERATURE", Springer, May-2011.

[10] Jeong-Yon Shim & Lei Xu, "MEDICAL DATA MINING MODEL FOR ORIENTAL MEDICINE VIA BYY BINARY INDEPENDENT FACTOR ANALYSIS", IEEE.P1-4, 2003.

[11] J.J.Tapia & E. Morett & E. E. Vallejo, "A CLUSTERING GENETIC ALGORITHM FOR GENOMIC DATA MINING", Foundations of Computational Intelligence, Studies in Computational Intelligence, Volume:204, 2009.

[12] J.Yanqing & H.Ying & J.Tran & P.Dews & A.Mansour & R.Michael Massanari, "MINING INFREQUENT CAUSAL ASSOCIATIONS IN ELECTRONIC HEALTH DATABASES", 11th IEEE International Conference on Data Mining Workshops, 2011.

[13] K.Sharmila & Dr.S.A.Vethamanickam, "SURVEY ON DATA MINING ALGORITHM AND ITS APPLICATION IN HEALTHCARE SECTOR USING HADOOP PLATFORM", International Journal of Emerging Technology and Advanced Engineering ISSN 2250-2459, Volume: 05, Issue: 01, January-2015.

[14] L.Chang & C.H.Chen, "APPLYING DECISION TREE AND NEURAL NETWORK TO INCREASE QUALITY OF DERMATOLOGIC DIAGNOSIS", Expert Systems with Applications- Elsevier, Volume: 36, pp. 4035-4041, 2009.

[15] Markus Brameier & Wolfgang Banzhaf, "A COMPARISON OF LINEAR GENETIC PROGRAMMING AND NEURAL NETWORKS IN MEDICAL DATA MINING", IEEE.p1-10, 2001.

[16] Michael Barnathan & Jingjing Zhang & Vasileios, "A WEB-ACCESSIBLE FRAMEWORK FOR THE AUTOMATED STORAGE AND TEXTURE ANALYSIS OF BIOMEDICAL IMAGES", IEEE. P1-3. 2008.

[17] O.Er & N. Yumusakc & F. Temurtas, "CHEST DISEASES DIAGNOSIS USING ARTIFICIAL NEURAL NETWORKS", Expert Systems with Applications- Elsevier, Volume: 37, pp. 76487655, 2010.

[18] Ping-Hung Tang &, Ming-Hseng Tseng, "MEDICAL DATA MINING USING BGA AND RGA FOR WEIGHTING OF FEATURES IN FUZZY K-NN CLASSIFICATION", IEEE.P1-6, July-2009.

[19] Pradnya P. Sondwale, "OVERVIEW OF PREDICTIVE AND DESCRIPTIVE DATA MINING TECHNIQUES", International Journal of Advanced Research in Computer Science and Software Engineering, Volume: 05 Issue: 04, April-2015.

[20] Prakash Mahindrakar & Dr. M. Hanumanthappa, "DATA MINING IN HEALTHCARE: A SURVEY OF TECHNIQUES AND ALGORITHMS WITH ITS LIMITATIONS AND CHALLENGES", Prakash Mahindrakar et al Int. Journal of Engineering Research and Applications: 2248-9622, pp.937-941, Volume: 03 Issue 06, Nov-Dec 2013.

[21] Ranjit Abraham & Jay B.Simha &Iyengar, "A COMPARATIVE ANALYSIS OF DISCRETIZATION METHODS FOR MEDICAL DATAMINING WITH NAÏVE BAYESIAN CLASSIFIER", IEEE. P1-2, 2006.

[22] R.Karthiyayini & J.Jayaprakash, "ASSOCIATION TECHNIQUE ON PREDICTION OF CHRONIC DISEASES USING APRIORI ALGORITHM", International Journal of Innovative Research in Science, Engineering and Technology, Volume: 04, Special Issue 06, May 2015.

[23] Sarojini Balakrishnan & Ramaraj Narayanaswamy, "FEATURE SELECTION USING FCBF IN TYPE II DIABETES DATABASES", Special Issue of the International Journal of the Computer, the Internet and Management, Volume: 17 No. SP1, March-2009.

[24] Sheetal L. Patil, "SURVEY OF DATA MINING TECHNIQUES IN HEALTHCARE", International Research Journal of Innovative Engineering, Volume: 01 Issue: 09, September-2015.

[25] Syed Zahid Hassan & Brijesh Verma, "A HYBRID DATA MINING APPROACH FOR KNOWLEDGE EXTRACTION AND CLASSIFICATION IN MEDICAL DATABASES". IEEE. P1-6, 2007.

[26] S. Soni & O. P. Vyas, "USING ASSOCIATIVE CLASSIFIERS FOR PREDICTIVE ANALYSIS IN HEALTH CARE DATA MINING", International Journal of Computer Applications, Volume: 04, No: 05, July-2010.

[27] Tsang-Hsiang Cheng & Chih-Ping Wei & Vincent S. Tseng, "FEATURE SELECTION FOR MEDICAL DATA MINING: COMPARISONS OF EXPERT JUDGMENT AND AUTOMATIC APPROACHES", IEEE. P1-6, 2006.

[28] U.Fayyad, G.Piatetsky-Shapiro and P.Smyth, "THE KDD PROCESS OF EXTRACTING USEFUL KNOWLEDGE FORM VOLUMES OF DATA", Communications of the ACM, pp. 27-34 Volume: 39, No: 11, November-1996.

[29] Usama Fayyad & Gregory Piatetsky & Padhraic Smyth, "Knowledge Discovery and Data Mining: Towards a Unifying Framework" KDD-96 Proceedings, 1996.

[30] Weimin Xue & Yanan Sun & Yuchang Lu, "RESEARCH AND APPLICATION OF DATA MINING IN TRADITIONALCHINESE MEDICAL CLINIC DIAGNOSIS", IEEE.p1-4, 2006.

[31] W.L.Zuoa & Z.Y.Wanga & T.Liua & H.L.Chenc, "EFFECTIVE DETECTION OF PARKINSON'S DISEASE USING AN ADAPTIVE FUZZY K-NEAREST NEIGHBOR APPROACH", Biomedical Signal Processing and Control, Elsevier, pp. 364373, 2013.

[32] Yanwei Xing & Jie Wang & Zhihong Zhao & Yonghong Gao, "COMBINATION DATA MINING METHODS WITH NEW MEDICAL DATA TO PREDICTING OUTCOME OF CORONARY HEART DISEASE", International Conference on Convergence Information Technology, 2007.

AUTHORS

Sheenal Patel received her B.C.A. and M.C.A. degree from Dharmsinh Desai University, Nadiad, Gujarat, India in 2012 and 2014 respectively. She is presently working as an Assistant Professor at Smt. Chandaben Mohanbhai Patel Institute of Computer Applications, Charusat University, Changa, Gujarat, India since 2014. Her research area include knowledge processing, data mining.

Hardik Patel has received M.C.A. degree from Smt. Chandaben Mohanbhai Patel Institute of Computer Applications, Charusat University, Changa, Gujarat, India in 2014 and B.C.A degree from M.B.Patel Science College, Anand, Gujarat, India. Now he is an Assistant Professor at Charotar Institute of Computer Applications – Changa, India. His research areas include data mining, cloud computing, content-based image and video analysis.