A Review of Data Mining Techniques for Health Care Analytics

[1]M.Ramesh,[2] M.Ramla,
[1]Assistant Professor, [2]Assistant Professor,
[1]Department of Computer Science, SRM Institute of Science and Technology, Chennai, India.
[2]Department of Computer Applications, SRM Institute of Science and Technology, Chennai, India.

Abstract— We now live in an era of complex data explosion with unprecedented flow. Drowning with data and Starving for Knowledge,has become the scenario. Especially in health care sector, the data is rich and massively flooded. With the application of Data Mining techniques, novel, useful and actionable insights can be uncovered to bring win-win strategy for both doctors as well as patients. Medical Diagnosis is an intricate task that needs to be carried out precisely and efficiently. This paper throws light into the Data Mining techniques used for medical prognosis and diagnosis.

IndexTerms: Health care data analytics, Data Mining, medical prognosis, medical diagnosis, Clinical prediction.

I. INTRODUCTION

Health care is an interdisciplinary field that encompasses diverse disciplines like Databases, Data Mining, Information Retrieval, medical researchers and computer scientists. It is extremely important to bring together all these and devise a novel data analytics tool to harness the power of massive flux of rich health care data.

Sources of Data

Approximately 500 Peta Byte of health care data is in existence today and the number is expected to skyrocket in the next seven years [1]. The ever increasing volume of data comes from variety of sources EHR, Biomedical Images, Biomedical Signals, Genome data, Clinical text data, Social media dataand Body Area Network (BAN) data.

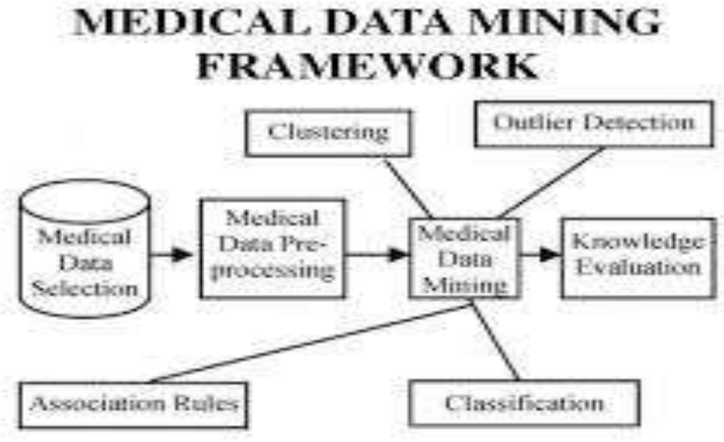
NEED FOR ANALYTICS IN HEALTHCARE

- Providing Patient centric services
- Detecting the spread of diseases earlier
- Monitoring the Quality of hospitals and preventing the hospital errors
- Improving the treatment methods without circumventing the treatments for patients
- Adverse effects of Drug Events.
- Fraudulence in credit card and Insurance claims.
- Predicting the risk of readmission of patients
- Predicting the length of stay of patients in hospital

CLINICAL PREDICTION

Clinical Prediction is an important branch of health care data analytics. It combines medical signs and symptoms in predicting the probability of a specific disease with risk assessment. The paper is organized by presenting a brief synopsis of the Data mining Techniques and then followed by the spectrum of clinical prediction research carried out by researchers.

Data mining is an assortment of algorithmic techniques to extract instructive patterns from raw data [4]. Data mining is the vital part of the KDD process and it can be broadly classified as Descriptive Data mining and Predictive Data mining. The KDD process consists of Selection, Pre-processing, Transformation, Data Mining and Interpretation/Evaluation. The various Data Mining techniques include Classification, Clustering and Association.



Classification

Classification is a two-step process consisting of trainingand testing. The first step, training, builds a classificationmodel, consisting of classifying rules, by analyzing trainingdata containing class labels. (An Example of a classification rule is "IF LungCancerFamilyHistory=yes AND Smoking=yes THEN CT_Scan=required").

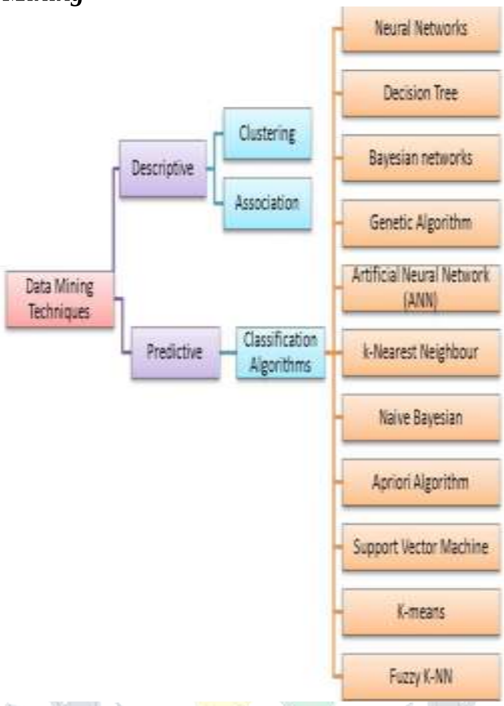
Clustering

Clustering is defined as unsupervised learning that occursby observing only independent variables[5]. (unlike supervisedlearning analyzing both independent variables and adependent variable).Clustering is used to group objects (records) into aspecific number of clusters so that the objects within acluster have very high similarity and objects from differentclusters have very low similarity. Similarities between twoobjects are measured using their attribute values.

Association

Association rule (or frequent item set/pattern) mining ina database (sometimes known as market basket analysis in asales transaction database), is often employed to discovercustomers' hidden sales patterns or relationships amongitems purchased. They can also be applied in medical data to identify the symptoms that goes together.

Different Techniques Used for Medical Data Mining



Challenges in Health Care Analytics

Health care data is highly sensitive and it reveals compromising information about individuals. Respecting the privacy of the patients, the health care data is often forbidden for access under HIPPAA act. It requires access to data by privacy preserving and data anonymity. The health care data is compromised in terms of its quality, periodicity and coverage. Good quality data is crucial for effective evidence based medicine. Effective concerns to bridge the gap in the data is a challenge

II. Review of Related Works

A good number of research has been carried out on diagnosis of different diseases. The aim of this paper is to have a keen observation of the various previous researches done in medical diagnosis.

In 2011, JyothiSoni et.al.[6] worked with an automatic medical diagnosis system for Cardio Vasculor Disease (CVD). Three different supervised machine learning algorithms, Naïve Bayes, k-NN, Decision List were used for data analysis. For predicting heart attack, significantly 15 attributes were identified like sex, chest pain type, blood sugar, Exang, Restecg, Serum Cholestrol, age, oldpeak etc. They concluded with the comment that Decision Tree outperforms Bayesian Classification

Genetic Algorithms was used to get the optimal subset of attributes to further improve the accuracy of decision tree and Bayesian classification.

Mohamed Abouzahra et al. [7] have implemented a model to integrate data from EHR to improve clinical decision making for Inflammatory Bowel Disease. The success rate of the drug recommendation was not done and privacy and security was not treated properly.

SankaraNarayanan et al.[8] invented a model for Diabetic Prognosis using the Apriori and FPGrowth to generate association rules for the Diabetes Mellitus dataset. But Scalability, Accuracy was not guaranteed for voluminous health records.

S. Muthukaruppan et al.[9] presented a Particle swarm optimization (PSO) based fuzzy expert system for the diagnosis of Coronary artery disease (CAD) based on the Cleveland Heart Disease datasets. Decision Tree technique was used to unravel the attributes and it resulted in crisp if-then rules converted into fuzzy rule base.

George et al. used Support Vector Machines to detect agitation transition. Haitham and Alan have projected automative recognition of obstructive sleep apnea syndrome using SVM classifier.

Fei Wang et al. [17] considered the problem of medical prognosis based on patient similarities and expert feedback. Given a query patient under investigation, they retrieve a cohort of similar patients using Euclidean Distance measure and Local Spline Regression.

In 2013, Gotlieb et al. [16]have accessed the ability of a large corpus of electronic records to predict the individual diagnoses. Their work suggested that one can harness the wealth of population based information embedded in EHR for predictive tasks. They compute patient similarities using only minimal set of measures. Genome and medical images could provide improved point of care for the target patient.

M. Yassi, A. Yassi and M. Yaghoobi (2014) [12] presented a paper to distinguish the type of breast cancer. They used chaotic hierarchal cluster-based multispecies particle swarm optimization (CHCMSPO) to optimize the fuzzy system.

In 2015, Fei Wang et al. presented a Patient Similarity Framework (PSF) that unifies and extends existing supervised learning metrics. It incorporates both unsupervised and supervised to learn patient similarity metrics.

KaliaOrphanou et al. [10](2016) presented a Naïve Bayes classification model where the features are temporal association rules to diagnose coronary heart disease (CHD). Temporal pattern mining algorithm was applied to detect TARs within the relevant patient history.

Muhammad Saqlianet al. (2016) [13] proposed a risk model to predict 1-year or more survival for Heart Failure patients using multinominal Naïve Bayes. The results showed that Naïve Bayes achieved a higher accuracy and AUC than any other models.

In 2016, PhattharatSongthung and KunwadeeSripanidkulcha[14] used Naive Bayes and CHAID (Chi-squared Automatic Interaction Detector) Decision Tree classifiers to predict high risk individuals and compared their results to existing hand-computed diabetes risk scoring mechanisms. Risk factors like age, gender, BMI, Waist circumference, Hypertension and Family history were considered. With Coverage as the evaluation metrics, they concluded with Naïve Bayes as the reasonable choice for predicting diabetic risk.

III. CONCLUSION:

The study reveals the importance of research in the area of life threatening disease diagnosis. The selection of Data Mining approach depends on the nature of data sets. There is no single Data Mining technique that perform well for health care domain. It is best to combine the approaches together for enhanced accuracy of health care analytics.

IV. REFERENCES

[1]. Data Driven Analytics in Healthcare:Problems, Challenges and Future Directions, Fei WangResearch Staff MemberIBM T. J. Watson Research Center[2]. Health care Data analytics, Chandan
K.Reddy, CharuC.Aggarwal

[2]. [2nd International Symposium on Big Data and Cloud Computing (ISBCC'15) A Survey Of Big Data Analytics in Healthcare and GovernmentJ.Archenaa1 and E.A.Mary Anita]

[3]International Journal of Information Sciences and Techniques (IJIST) Vol.6, No.1/2, March 2016 SURVEY OF DATA MINING TECHNIQUES USED IN HEALTHCARE DOMAIN Sheenal Patel and Hardik Patel

[4] Data Mining in Healthcare and Biomedicine: A Surveyof the LiteratureIllhoiYoo& Patricia Alafaireet&Miroslav
Marinov&Keila Pena-Hernandez &RajithaGopidi&Jia-Fu Chang & Lei Hua

[5] Predictive DAta Mining for Medical Diagnosis: An overview of Heart Disease
Prediction, International Journal of Computer Applications, Vol.17-No.8, March 2011

[6] Mohamed Abouzahra, Kamran Sartipi, David Armstron, Joesph Tan, 'integrating Data from EHR to enhance clinical decision making: The Inflammatory Bowel Disease case" Proceedings of 27th International Symposium on Computer-Based Medical Systems, 2012

[7] Sankaranarayan.S and Padmanadaperumal.T, "Diabetic prognosis using Data Mining methods and techniques", Proceddings of ICICA, Coimbatore, India 2014

[8] S.Muthukarrupan "A hybrid particle swarm optimization based on fuzzy expert system for the diagnosis of coronary artery disease", Expert Systems with applications, Vol.39, Issue 14,Oct 2012 Pages 11657-11665

[9]Combining Naive Bayes Classifiers withTemporalAssociation Rules for Coronary Heart DiseaseDiagnosis, 2016 IEEE International Conference on Healthcare Informatics, KaliaOrphanouuet

[10] Medical Data Mining UsingDifferent Classification and ClusteringTechniques: A Critical Survey, 2016 Second International Conference on Computational
Intelligence & Communication Technology, RichaSharma et al.

[11]M. Yassi, A. Yassi and M. Yaghoobi, "Distinguishingand clustering breast cancer according to hierarchalstructures based on chaotic multispecies particle swarmoptimization", Iranian Conference on IntelligentSystems, pp 1-6, feb 2014.

[12] 2016 45th International Conference on Parallel Processing Workshops Identification of Heart Failure by Using UnstructuredData of Cardiac PatientsMuhammad Saqlain, Wahid Hussain, Nazar A. Saqib, Muazzam A. Khan College of Electrical and Mechanical Engineering (E& ME) National University of Sciences and Technology

[13] Improving type 2 Diabetes Mellitus Risk Prediction using Classification, PhattharatSongthung and KunwadeeSripanidkulchaiNational Electronics and Computer Technology Center (NECTEC), 2016 13 th International joint Conference on Computer Science and Software Engineerig (JCSSE)

[14]2013 International Conference on Green Computing, Communication and Conservation of Energy (ICGCE), A survey on mining techniques for Early Lung Cancer Disease.

[15] Gottlieb et al. "A method for inferring medical diagnoses from patient similarities" ,BMC Medicine, 2013.

[17] Fei Wang, Jianying Hu, Jimeng Sun, "Medical Prognosis based on Patient Similarity and Expert Feedback" , ICPR 2012