A Collective Study on Mining Health Records

Amisha Wankhede[1], Dimpal Adate[2], SnehaPawar[3,] Harshita[4], N.K. Patil[5]
[1]UG Student, Dept. of Computer Engineering, Savitribai Phule Pune University
[2]UG Student, Dept. of Computer Engineering, Savitribai Phule Pune University
[3]UG Student, Dept. of Computer Engineering, Savitribai Phule Pune University
[4]UG Student, Dept. of Computer Engineering, Savitribai Phule Pune University
[5]Assistant Professor, Dept. of Computer Engineering, Savitribai Phule Pune University

Abstract—**In this paper the study Overall health inspection is companion essential a part of care in several countries. Distinctive the participants in hazard are very important for early notice and preventive intervention. The fundamental challenge of learning a classification model for risk forecast lies within the unlabeled knowledge that establishes the bulk of the collected dataset. There's no ground truth for discriminating their states of health. Significantly, the unlabeled knowledge describes the contributors in health investigations whose health conditions will vary greatly from healthy to very-ill. In this paper, we tend to recommend a graph-based, semi-supervised learning algorithmic rule mentioned to as SHG-Health (Semi-supervised Heterogeneous Graph on Health) for risk predictions to categorize an increasingly developing scenario with the bulk of the information unlabeled Wide-ranging experiments supported each real health examination datasets and artificial datasets are achieved to indicate the effectiveness and strength of our procedure.**

Keywords—Health examination records, semi-supervised learning, Electronic Health Records

I. INTRODUCTION

HUGE amounts of Electronic Health Records (EHRs) collected over the years have provided a rich base for risk analysis and prediction An EHR contains digitally stored healthcare information about an individual, such as observations, laboratory tests, diagnostic reports, prescriptions, patient identifying knowledge , and allergies A special type of EHR is the Health Examination Records (HER) from yearly general health check-ups. For example, governments such as Australia, U.K., and Taiwan offer periodic geriatric health examinations as an integral part of their aged care programs. Since clinical care often has a specific problem in mind, at a point in time, only a confined and often small set of measures considered necessary are collected and stored in a person's EHR. By contrast, HERs are collected for regular surveillance and preventive purposes, covering a comprehensive set of general health measures , all collected at a point in time in a systematic way. Identifying participants at risk based on their current and past HERs is important for early warning and preventive intervention. By "risk", we mean unwanted outcomes such as mortality and morbidity. In this study we formulated the task of risk prediction as a multi-class classification problem using the Cause of Death (COD) information as labels, regarding the health-related death as the "greatest risk". The goal of risk prediction is to effectively categorize 1) whether a health examination participant is at risk, and if yes, 2) predict what the key associated disease category is. A good risk prediction model should be able to exclude low-risk situations and clearly identify the high-risk situations that are related to some specific diseases A fundamental challenge is the large quantity of unlabeled data .For example,92.6%ofthe102,258participantsin our geriatric health examination dataset do not have a COD label. The semantics of such "alive" cases can be different from generally healthy to seriously ill or anywhere in between. In other words, there is no truth shown for the "healthy" cases. If we simply treat this set of alive cases as the negative class, it would be a highly noisy majority class. On the other hand, if we take this large set as genuinely unlabeled, as opposed to cases with known labels removed, it would become a multi-class learning problem with large unlabeled data. Methods that consider unlabeled data are generally based on Semi-Supervised Learning (SSL) that learns from both data. Mining health examination information and learning methods that manage unlabeled health data.Data mining is described as "a process of nontrivialextraction of implicit, previously unknown andpotentially useful information from the data stored in adatabase" Healthcare databases have alarge amount of data but , there is a lack ofeffective analysis tools to discover the hiddeninformation. Adequate computer-based informationor decision support systems can help physicians. Efficient and precise implementation of an automated system needs a comparative study of various techniques. Here we present an overview of thepresent research being carried out using the DMtechniques for the diagnosis and prognosis of variousdiseases, highlighting critical issues and summarizing theapproaches in a set of learned lessons.

Data mining tools are used for decision making. Prediction and classification techniques are used in which classification technique predicts the unknown values with respect to generated model. An assortment of data mining techniques can be implanted to find associations and patterns in data, extract knowledge in the forms of rules and predict the value of the dependent variables. Common data mining techniques which are used in almost all the sectors are listed as: Naive Bayes, Decision Tree, Artificial neural network (ANN), Bagging algorithm, K- nearest neighborhood (KNN), Support vector machine (SVM) etc. Data mining is an important step of knowledge discovery in databases (KDD) which is an iterative process of data cleaning, integration of data, data selection, pattern recognition and data mining knowledge recognition. KDD and data mining are also used interchangeably. Data mining encompasses association, classification, clustering, statistical analysis and prediction.

Drug is any substance that when taken into a living organism may alter one or more of its processes. Drugs can provide temporary relief from unhealthy symptoms and/or permanently supply the body with essential substances the body can no longer make. Some drugs lead to an unhealthy dependency that has both physiological and behavioral roots. Drug addiction can cause serious, long-term consequences, including problems with physical and mental health, relationships, employment, and the law . Adolescence is typically a period of experimentation, irrespective of parenting skills and influence. However, the more likely threat to any teenager's health is the use of drugs such as alcohol and tobacco.The increased adoption of electronic health records (EHRs) has led to an unprecedented amount of patient health information stored in electronic format. However, the availability of overwhelmingly huge records hasalso raised concerns of information overload, with potential negative consequences on clinical work, such as errors of omission, delays, and overall patient safety. Current EHR systems often do not present this tremendous amount of patient data in a way that supports clinical workflow or cognitive reasoning. It is therefore imperative for patientcare to automatically go through the raw data points present in the records and detect timely and relevant information.Alarmingly, as the most chronically ill patients often have the largest datasets, their records are the most difficult to coherently present. As an example, for a prevalent chronic condition in our institution, patients with chronic kidney disease have 338 notes on average in their record (from all clinical settings) gathered across an average of 14 years, with several patients' records containing over 4000 notes. It is clear that during a regular medical visit, no practitioner can read hundreds of clinical notes. Fortunately, electronic storage of this health information provides an opportunity for EHR systems to "aid cognitionthrough aggregation, trending, contextual relevance, minimizing superfluous data." Currently available commercial EHR systems, however, inadequately address this need, sometimes providing organization of data but lacking in information synthesis.8 Some vendor EHR dashboards display problem lists that aggregate billing codes butthese are low in actionable knowledge.

II. LITERATURE SURVEY

In this section of paper some important works are being analyzed to employ the feature of health mining as follows:

[1] Dr. D.P. Shukla and ShamsherBahadur Patel proposes algorithm for data classification, clustering, regression, association and rule mining, (CART) Classification and Regression tree. Limitations observed are voluminous data produced cannot be managed and scope improving quality of prediction diagnosis and disease classification.

[2] DivyaTomar and SonaliAgarwalpresent the methodology of AprioriAlgorithm, clustering, regression. Demerits to this concept are that it cannot maintain relevant medical data, difficult to acquire precise health care data, data is complex, does not give consistent results. Future scope of this concept is that we can use hybridization or integrate Data Mining technology such as fusion of different classifiers,fusion of clustering with classification, association for better performance.

[3]SupritKaur and Dr. R.K. Bawa introduce Knowledge discovery and database (KDD), Medical diagnosis and prognosis. The disadvantages of this paper are that the applying data mining in the medical field is incredibly challenging mission due to idiosyncrasies of medical profession. The future scope is that it involves amalgamation of various specified algorithm to augment the accuracy so that the diagnosis can develop into more accurate data sets. It mainly focuses on knowing the indulgement of youth in drugs.

[4] RimmaPivovarov and NoemiElhadadgive the knowledge about Electronic health records(EHR) summarization. The limitations to this are: Chronically ill patients often have largest datasets, their records are most difficult to present. The future of this piece of knowledge is rich, complex and essential health data of millions of patients, the informatics community has new opportunity to tackle challenges of interpreting a mounting wealth of health information.

[5] Ling Chen, Xue LI introduces Personal Health Indexing and

Geriatric Medical Examination. the Demerits of this methodology is to optimize problems that find optimal of labels as health score based on medical records that are infrequent, incomplete and sparse. Evaluation of Health care status of a person from cradle-to-grave is becoming possible.

[6] N Collier, S Doan and A. Kawazoe have proposed Ontology based text mining, Naive entity recognition,location detection and event recognition. Massive volume of data need to interpret information as soon as possible in the outbreak cycle when reliable fact tend to be scarce are its disadvantages. Extending coverage to new languages and public health threats are the demerits to this concept.

[7] SE. Brosette, AP. Sprague and JM. Hardin experiments with Data Mining Surveillance system to analyze pseudomonas aeruginosa. They couldn't handle too much voluminous data. The further experiments will deal with public health and intensive care unit infection control data, utilizing prospective clinical studies.

[8] Bath, P.A. deals with Data Mining, Artificial neural networks, Machine learning, Decision tree, Rule based evolutionary, Genetic Algorithm. Results may not be accurate to this methodology. They will be widely recognized as complementary to traditional methods of analyses data in health and medicine.

[9] YC. Huang proposed Association rule correlations, empirical data and high frequency patterns are identified. Apriori data scan results in low efficiency, adds complexity by involving two differentdatabases to find correlations between multiple diseases and abnormal result in health examination. Association rules can serve valuable reference guides for health management and health care personnel.

[10] M.S. Viveros, J.P. Nearhos and M.J. Rothman introduced association rules and neural segmentation. By applying implementations on self organizing maps we found that there is no correct number of segments. Data mining algorithms can be used on large, real customer data with reasonable execution time.

III. CONCLUSION

Most of the time due to excess commitments to other things we are not able to give much priority to our health. So this paper analysis many papers which are about to examine the health records of the people to identify any worst case scenarios ahead or not. By collective study of all, this paper feels graph approach will be more affective due to its hierarchical analysis of the health record. And due to this time complexity can be manage to some extent on increase of input data.

References

[1] Dr. D.P. Shukla and ShamsherBahadur. A Literature review in health informatics using Data mining technology; 2014, IJSHRE

[2] DivyaTomar and SonaliAgarwal. A survey on data mining approaches for healthcare; 2013, IEEE

[3] SupritKaur and Dr.R.K. Bawa.Future trends of Data Mining in predicting the various diseases in medical healthcare system; 2015, IJEIC.

[4] RimmaPivovarov and NoemieElhadad automated methods for the summarization of electronic health records; 2012, IRJET.

[5] Ling Chen and Xue Li. Personal health indexing based on medical examination: A Data Mining approach; 2014, IEEE.

[6] N Collier, S Doan, A. Kawazoe. Bio Caster: detecting public health rumors with a health based text mining system; 2008, Bioinfomatics.

[7] SE. Brosette, AP. Sprague. Association rules and Data Mining in hospital infection control and public health surveliance; 2011, IEEE.

[8] Bath, P.A.Data Mining health and medical information; 2004, IEEE.

[9] YC Huang. Mining Association rule between abnormal health examination results and outpatient medical records; 2010, IJEIC

[10] M.S. Vineros, J.P. Nearhos. Applying Data mining technique to health insurance information system; 2015, IRJET.