International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 12 | Dec -2016 www.irjet.net p-ISSN: 2395-0072

A Review on health care examination records using data mining

Manoranjani A. Kulkarni[1],Prof.Chaitanya S. Kulkarni[2]

[1]ME Student ,Dept of Computer ,JSPM NTC,PUNE,India
[2]Professor,Dept of computer Engineering ,JSPM NTC,Pune,India

---

Abstract - *The general medical examination is a typical type of preventive medication including visits to a general expert by well feeling adults on a regular basis. Making out the ones taking part at risk is important for early suggestions and precautions coming between groups. The big challenge of learning the design for risk of unhealthy life in future lies in the unlabeled data which is a very integral part of the dataset which consist of the person's data who is perfectly healthy and whose condition varies from healthy to ill. In this paper, they propose a graph-based, semi-supervised learning algorithm called SHG-Health (Semi-supervised Heterogeneous Graph on Health) for risk predictions of what will take place in the future to put in order a by degrees undergoing growth place, position with the greater number or part of the facts without mark, name. Here, they will focus mainly on unlabeled data so that system will work for both undiagnosed patient and the healthy one. With this system, people will be getting intimate precaution before even dealing with a disease. Hence, this system will lead to a healthy life.*

Key Words: Health examination records, semi-supervised learning, heterogeneous graph extraction.

1.INTRODUCTION

As the technology have proliferated in recent years, people are shifting to a new era where our life has started revolving around technology. These things have made human life much simpler. One of the major area where the technology have proved to be more useful is Medical. Our purpose is to make the health care system more reliable. Patient's Health Record have been saved in a system for many years. An Electronic Health Records (EHR) stores all the details of patients including physical details, allergies, primordial diseases and the diseases the person have dealt so far. For doing so, a health examination programs have been conducted in primitive years and has been stored in Health Examination Records (HER). By contrast, HERs arecollected for regular surveillance and preventive purposes,
Covering a comprehensive set of general health measures. The EHR received its first real validation in an Institute of Medicine's (IOM) report in 1991 entitled "The Computer-Based Patient record: An Essential Technology for Health care"[3]. IOM drove home the idea that the EHR is needed to transform the health system to improve quality and enhance safety. The EHR becomes a tool through which the family medical office can transform practice to meet its need and need of the patient. Improved work processes and access to data make the practice of medicine more effective for doctors and their staff. Decision support and automated reminders help the practice deliver safer and higher quality care to patients and the community. The EHR is about quality, safety, and productivity. It is an extraordinary apparatus for doctors, however can't guarantee these virtues in isolation. Achieving the true benefits of EHR systems requires the transformation of practices, based on quality improvement methodologies, system and team based care, and evidence-based medicine. We face a huge challenge, when it comes for retrieving a patient's record from billions of records.

For creating the mentioned model, we need to focus on unlabeled data which can be mostly done by Semi Supervised Learning. Semi-Supervised Learning is a situation in which in your training data some of the samples are not labeled. The semi-supervised estimators can make utilization of this extra unlabeled data to better capture the state of the underlying data distribution and generalize better to new samples[5]. These algorithms can perform well when we have a very small amount of labeled points and a large amount of unlabeled points. But the real challenge in EHR is its heterogeneity. Therefore, our system proposes a semi-supervised heterogeneous graph based algorithm called SHG-Health as a predictive model for risk calculation. To handle heterogeneity, it explores a Heterogeneous graph based on Health Examination Records called **HeteroHER** graph, where examination items in different categories are modeled as different types of nodes and their temporal relationships as links[2]. To tackle large unlabeled data, SHG-Health features a semi-supervised learning method that utilizes both labeled and unlabeled instances. In addition, it is able to learn an additional K +1"unknown" class for the participants who do not belong to the K known high-risk disease classes.

1.1 OVERVIEW OF SHG-HEALTH ALGORITHM

GHE Dataset:- It de-identified database all private data, such as name ,contact details ,birth dates removed.The dataset has 230 attributes ,containing 264,424 check-ups of 102,258 participants aged 65 or above[1].

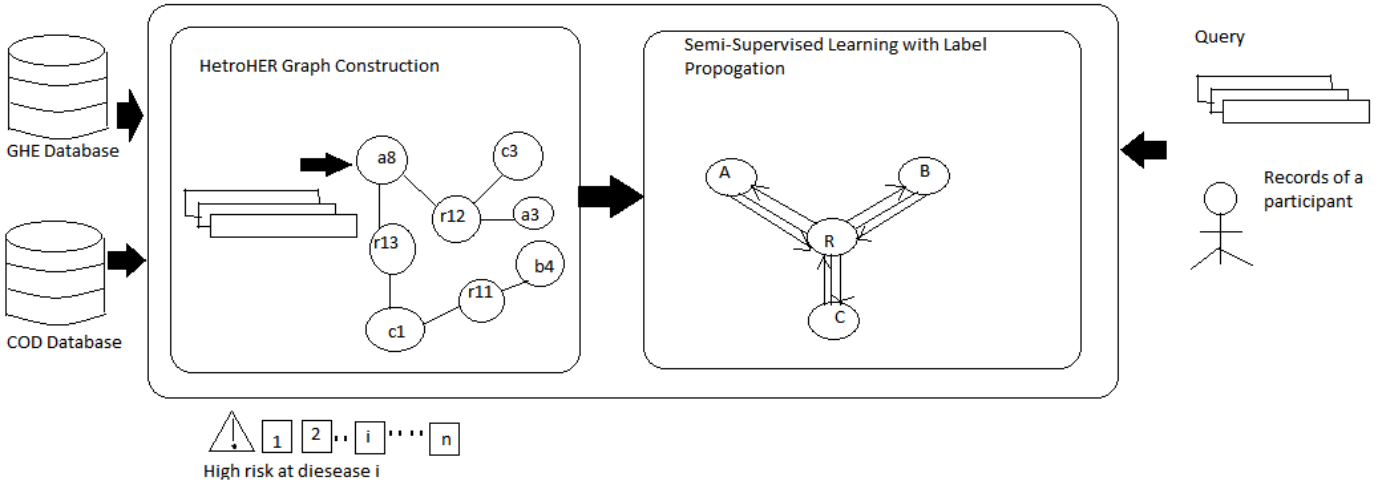

Fig 1. An overview of the SHG-Health Algorithm

COD Dataset:- The GHE dataset was linked to the Taiwan National Death Registry system using participants' identification numbers and then encrypted to provide de-identified secondary data maintained by the Department of Health of the Taipei City Government. We called this linked subset of data the Cause of Death dataset[1].

Take set of records as an input and construct a heterogeneous graph with it. Following are the steps for graph construction.

a. Binarization: Our first step is to convert the value of each node in a record into 1 or 0 which indicates the presence of discretized value.
b. Node Insertion: For node insertion, we consider only those nodes whose result is abnormal with binary value of 1.
c. Node typing: Every node is then categorized according to their examinations. We have considered mainly three categories: Physical Test(A), Mental Test(B) and profile(C).
d. Link Insertion:These nodes are then linked to other nodes which are not a part of patient's record (eg. Attributes of a disease). After linking the nodes, weight of the links is calculated based on the assumption that the newer a record the more important it is in terms of risk prediction.

2.RELATED WORK

M. S. Mohktar[1],The objective of this study was to develop and validate a classification algorithm for the early identification of patients, with a background of chronic obstructive pulmonary disease (COPD), who appear to at high risk of an imminent exacerbation event. The algorithm attempts to predict the patient's condition one day in advance, based on a comparison of their current physiological measurements against distribution of their measurements over the previous month. M. Zhao[2] This paper has introduced an effective semi-supervised learning algorithm, which is based on a newly proposed graph that can represent the data manifold structure in a more compact way. Also, this model has proposed CGSSL algorithm for Medical Diagnosis.This paper was only implemented for neurological disorders among the elderly. There wasn't any mention of unlabeled class. E. Kontio[3] This paper proposed that, by applying language technology to electronic patient documents it is possible to accurately predict value of the acuity scores of the coming day based on the previous day´ s assigned scores and nursing notes.This paper do not consider the issue of unlabelled data. They either have expert-defined low-risk or control classes or simply treat non-positive cases as negative.C. Y. Wu[4]To determine whether cognitive impairment assesse at annual geriatric health examinations is associated with increased mortality in elderly. This paper considers small set of measures that are necessary and are collected and stored in a person's EHR.Y. Zhao[5]The goal is to 1) find groups of samples corresponding to different phenotypes (such as disease or normal), and 2) for each group of samples, find the representative expression pattern. This paper is limited to only some related subclasses of labelled data.M. F. Ghalwash[6] Leveraging temporal observations to predict a patient's health state at a future period is very challenging task. Providing such a prediction early and accurately allows for designing a more successful treatment that starts before a disease completely develops.There is no get onto land truth

for differentiating their states of being healthy. Y. Sun[7]In this paper, we address a new clustering problem to detect net-clusters on special heterogeneous network with star network schema.The methods in this paper were designed for a multi-class semi-supervised learning problem with predefined classes, and thus have no mechanism for handling the"unknown" class.H. Huang[8]This paper has present enhanced semi-supervised local fisher discriminant analysis method for dimensionality reduction, which exploits both statistically uncorrelated and parameter-free charactristics .This paper does not consider an "unknown" class and they all have predefined instances for all classes, either by experts or via other mechanisms. In addition, all the graph based SSL methods used in this paper has homogeneous graphs.Y. Sun[10] In this paper, we address new clustering problem to detect net-clusters on special heterogeneous network with star network schema.The methods in this paper were designed for a multi-class semi-supervised learning problem with predefined classes, and thus have no mechanism for handling the "unknown" class. Inspired.

3.MOTIVATION:

The purpose of this project is to develop a system where a person can get his/her health risk based on the previous health conditions. This will help people to take precaution before even getting the disease.

4.PROBLEM STATEMENT:

The problem in current state of art unlabeled data gives a detailed account of the ones taking part in being in healthy examination whose being healthy conditions can different greatly from healthy to very-ill. There is no get onto land truth for differentiating their states of being healthy.

5.EXISTING ALGORITHM:

Input: a set of health examination records of n participants

S, the corresponding encoded labels Y

Output: optimized F as the computed soft label

Step 1: W<-Graph construction from S

Step 2: Calculate the normalized weights for I, j = 1,……m by:
$\hat{W}_{ij} = D_{ij}^{-1/2} W_{ij} D_{ji}^{-1/2}$

Where W_{ij} is the weight between the two nodes. It can be calculated as:

$g(t)= (t-s+1)/l$

Where t is the time window of interest, and s is the starting time of the time window.

Step 3: Initialize F_i uniformly amongst type i nodes for i = 1,……m.

$J(F) = \sum_{ij}^{m} \gamma_{ij} \sum_{p}^{ni} \sum_{q}^{nj} \hat{W}_{ij,pq} ||F_{ip}-F_{jq}||^2_F + \sum_{i}^{m} \sum_{p}^{ni} \mu_{ip} \hat{d}_{ip} ||F_{ip} - Y_{ip}||_F^2$ Where

γ_{ij} = ½ z_j , if i and j are same type of nodes
Z_j ~otherwise

Also, $\hat{d}_{ip} = \sum_{j}^{m} \sum_{p}^{n} z_j \hat{D}_{ijpp}$ And $\mu_{ip}>0$

Step 4: make time t = 1

Step 5: Repeat

Step 6: Update Fi for i = 1,…..m by:

$F_i(t+1) = I_{\alpha i} \sum_{j}^{m} z_j P_{ij} F_j(t) + I_\beta Y_i$

Where $P_{ij} = \hat{D}_{ij} \hat{W}_{ij}$

Step 7: Increment t by 1 i.e. t=t+1

Step 8: Repeat the above step until

$\lim_{t\to\infty} F(t) = (I - \hat{P})^{-1} I_\beta Y$

Step 9: Return F

6.PROPOSED SYSTEM:

In this proposed system,I am generating a centralizes system, in this system doctor and patient both first need to do registration. The basic information for registration will be name ,address ,phone no, email id, gender etc. After registration either patient or doctor can feel symptoms to find risk. When patient come for any test like HB ,Sugar ,Urine test etc., after test we send its report to patient account. With the help of data extraction system will generate report, prediction and also display graph as well as provide prescription.

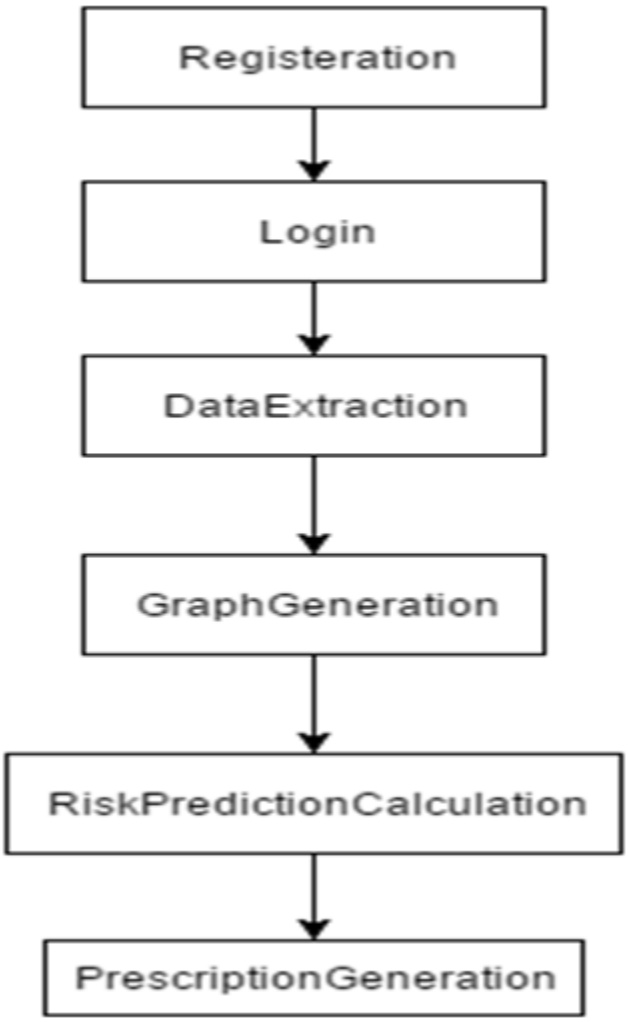

Fig 2. An overview of the proposed System

Data extraction is where data is analyzed and crawled through to retrieve relevant information from data sources in specific pattern. With the help of prediction risk will be generated the will be generated through graph. Another functionality of the system is that, for regular check up of patient a system well send direct notification on patient account.

7.CONCLUSION

In this paper, 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 classification model for risk forecast lies within the unlabeled knowledge that establishes the bulk of collected dataset. Theres 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, author tend to recommend a graph-based, semi-supervised learning algorithmic rule mentioned to as SHG-Health for risk predictions to categorize 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 procedure. Associate economical repetitive algorithmic rule is projected and therefore the proof of conjunction is given. In this system, Health reocrds are represented as graph so that is useful for developing abnormal results.

Future work will be generation of prescription, sending reports to the patient on personal account and for regular check up of patient a system well send direct notification on patient account.

8.REFERENCES

1. Ling chen,xue Li,"Mining health examination records-a graph based approach"IEEE Transaction on Knowledge and Data Engineering,pp1041-4347,2016
2. M. F. Ghalwash, V. Radosavljevic, and Z. Obradovic, "Extractionof interpretable multivariate patterns for early diagnostics," IEEEInternational Conference on Data Mining, pp. 201–210, 2013.
3. M. S. Mohktar, S. J. Redmond, N. C. Antoniades, P. D. Rochford, J. J. Pretto, J. Basilakis, N. H. Lovell, and C. F. McDonald, "Predicting the risk of exacerbation in patients with chronic obstructive pulmonary disease using home telehealth measurement data," Artificial Intelligence in Medicine, vol. 63, no. 1, pp. 51–59, 2015.
4. M. Zhao, R. H. M. Chan, T. W. S. Chow, and P. Tang, "Compact Graph based Semi-Supervised Learning for Medical Diagnosis in Alzheimer's Disease," IEEE Signal Processing Letters, vol. 21, no. 10, pp. 1192–1196, 2014.
5. P. Yang, X. L. Li, J. P. Mei, C. K. Kwoh, and S. K. Ng, "Positive unlabeled learning for disease gene identification," Bioinformatics, vol. 28, no. 20, pp. 2640–2647, 2012.
6. E. Kontio, A. Airola, T. Pahikkala, H. Lundgren-Laine, K. Junttila, H. Korvenranta, T. Salakoski, and S. Salanter¨a, "Predicting patient acuity from electronic patient records." Journal of Biomedical Informatics, vol. 51, pp. 8–13, 2014.
7. H. Huang, J. Li, and J. Liu, "Gene expression data classification based on improved semi-supervised local Fisher discriminant analysis," Expert Systems with Applications, vol. 39, no. 3, pp. 2314– 2320, 2012.
8. Y. Sun, Y. Yu, and J. Han, "Ranking-based clustering of heterogeneous information networks with star network schema," in Proceedings of the 15th ACM

SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2009, pp. 797–806.

9. J. Kim and H. Shin, "Breast cancer survivability prediction using labeled, unlabeled, and pseudo-labeled patient data," Journal of the American Medical Informatics Association : JAMIA, vol. 20, no. 4, pp. 613–618, 2013.
10. C. Y. Wu, Y. C. Chou, N. Huang, Y. J. Chou, H. Y. Hu, and C. P. Li, "Cognitive impairment assessed at annual geriatric health examinations predict mortality among the elderly," Preventive Medicine, vol. 67, pp. 28–34, 2014.
11. Y. Zhao, G. Wang, X. Zhang, J. X. Yu, and Z. Wang, "Learning phenotype structure using sequence model," IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 3, pp. 667–681, 2014.