To Develop Healthcare Approach using Clustering

Manoranjani A. Kulkarni
RSSOER, JSPM,
NTC, Pune

Chaitanya S. Kulkarni
RSSOER, JSPM,
NTC, Pune

ABSTRACT
Health care domain have attracted considerable amount of research fields. One of the field that has a drastic focus on health care domain is data mining. Mainly health care system focuses on some data mining theories like classification, clustering etc. The backbone in the domain of data mining is the data itself. For any field that is related to data mining, the data should be reliable and huge. System is working on patient's medical data i.e. electronic health record. A large amount of diagnosed patient's medical test data is stored electronically on a local machine. The aim is to provide such an unwavering service to the patient so that the patient should have complete knowledge of their disease before going for diagnosis. A system can predict the disease by considering few parameters of the patient's test. Patient's disease can be easily detected without wasting days for waiting for their test's results. This prediction system is implemented by using classification algorithm i.e. semi-supervised heterogeneous graph-based algorithm. The Proposed system should be compatible to provide not only the prediction but should also calculate their prescription, dosage, and health check-up status. Proposed system does not only benefit the patients but the doctors. k-means algorithm is implemented for clustering the patient who are at risk. When the clustering of risked patient is formed, doctor will also have the facility to notify the patient about their risk via e-mail. This approach will help us to save time in the diagnosis process and will make health care system a well-grounded one.

General Terms
Heart blockage classification, thyroid classification.

Keywords
Health examination records, semi-supervised learning, heterogeneous graph extraction.

1. INTRODUCTION
Health care domain is playing a vital role in technologies now a day. People are willing to step forward for more healthier life. This domain has attracted many researchers to make it more reliable. The main motivation is to make the decision of a person's health fast and dependable. 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 are collected for regular surveillance and preventive purposes, covering a comprehensive set of general health measures. Computing on an Electronic health record is a critical task. The huge challenge, when it comes for retrieving a patient's record from billions of records. For creating the mentioned model, there is need to focus on unlabeled data which can be mostly done by Semi Supervised Learning. Semi-Supervised Learning is used when data contain both unknown and known labels. The semi-supervised learners uses the additional unlabeled data to shape the data distribution and to generalize them better. But the real challenge in EHR is its heterogeneity. Therefore, proposed 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. To tackle large unlabeled data, SHG-Health features a semi-supervised learning method that utilizes both labeled and unlabeled instances. In addition, it can learn an additional K +1 "unknown" class for the participants who do not belong to the K known high-risk disease classes.

2. RELATED WORK
Breast cancer survivability prediction[2]
Background studies of breast cancer survivability have been assisted by machine learning algorithms, which can predict the survival of a particular patient based on medical history of other diagnosed patients. This paper has proposed a machine learning approach by Semi Supervised Learning. However, SSL should have more labeled data for the better result because it is a learning algorithm guided by information contained in the labelled data, like other machine learning algorithms. To compensate for the lack of labeled data, therefore, SSL Co-training was proposed in this paper.

Cognitive impairment assessed[3]
To determine whether cognitive impairment assessed at annual geriatric health examinations is associated with increased mortality in the elderly. This troop study was based on data obtained from the government-sponsored Annual Geriatric Health Examination Program for the elderly in Taipei City between 2006 and 2010. The study sample consisted of 77,541 community-dwelling Taipei citizens aged 65 years or older.

Semi-Supervised Learning for Diagnosis in Alzheimer's Disease[5]
This paper has introduced a graph based semi-supervised learning algorithm for Medical Diagnosis by using partly labeled samples and large amount of unlabeled samples. The newly proposed graph can represent the data manifold structure in a more compact way. Therefore, proposed CGSSL algorithm for Medical Diagnosis.

Extraction of Interpretable Multivariate Patterns for Early Diagnostics[6]
This paper has proposed interpretable patterns for early classification (IPED). The IPED method starts with transforming the multivariate time series data into a binary matrix representation over the span of all extracted shapelets from all the dimensions of the time series. IPED method addresses three issues in the state-of the- art MSD method. To start with, the parts of the multivariate shapelet don't have the limitations of a similar beginning time point and are not required to be of a similar length.

Improved semi-supervised local Fisher discriminant analysis[7]
This paper has presented an enhanced semi-supervised local fisher discriminant analysis method for dimensionality reduction, which exploits both statistically uncorrelated and parameter-free characteristics. iSELF can preserve the global structure of unlabeled samples in addition to separating labeled samples in different classes from each other, and so it efficiently extracts the discriminant information in the low dimensional embedding space and addresses the semi-supervised learning problem for gene expression classification.

Phenotype Structure Using Sequence Model[8]
This paper has modelled the phenotype structure discovery problem from a sequence perspective not the same as alternate strategies, the proposed g-successions show utilizes the requested quality expression esteems as the discriminative marks. In the FINDER algorithm, a novel sequence dissimilarity measurement and a cross projection approach enable to try exploring candidate phenotype structures in a quality-guaranteed way.

Positive-unlabeled learning[9]
This work has proposed a novel PU learning approach PUDI for disease gene prediction. They introduced a new feature selection method to identify the discriminating features and performed a further partitioning of the unlabeled set U into multiple training sets for a more refined treatment of U to assemble the end classifier.

Predicting patient acuity from electronic patient records[10]
By applying language innovation to patient archive it is conceivable to consummate predict the estimation of the keenness scores of the coming day in light of the earlier day' s doled out scores and nursing notes. The consequences of this review affirm that it is conceivable to utilize electronic literary nursing notes and already doled out keenness scores to predict a patient's sharpness for the following day through the use of machine learning strategies.

3. EXISTING SYSTEM

The idea was to design an automatic system which will predict the result of electronic health examination records. As a discussed in section 2, there are many approaches for classification of diseases. Previous system focuses on the following steps:

1.Graph Construction: Before training the dataset, a graph is constructed by converting each value of all the attributes into nodes having binary values. These nodes are then categorized into their types i.e. labeled and unlabeled. Also, weight between each node is calculated by:

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

2.Risk Calculation: After the graph is constructed, Risk calculation is done. For risk prediction, There is need to implement a classification algorithm by training, testing and cross-validation. This prediction variable is named as soft variable and is calculated by:

$$F_i(t+1) = I_{\alpha i} \sum_j^m z_j P_{ij} F_j(t) + I_\beta Y_i \quad (2)$$

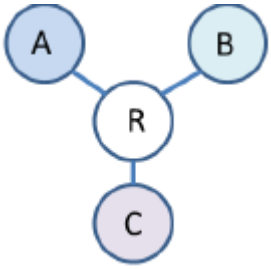

Fig 1:- Schema(1)

Given a set of health examination records of participants, There is need to first classify every record to labeled and unlabeled. The classification will be done based on the result of three heath examination vise Physical Test(A), Mental Test(B) and Profile(C). The goal is to predict for unlabeled as well as labeled class. Also, to provide the prescription and notification for heath conditions.

4. PROPOSED SYSTEM

In a proposed system, the new algorithm is introduce for identifying the patient's level of health risk with the resulted soft label. By identifying the health risk of patient, System can generate the prescription. Prescription is generated only for those patients whose health risk is above the threshold and who needs serious attention. Along with the prescribed drugs, Proposed system also providing the detailed description of them and the dosage is calculated according to their value of soft label. Along with that, clustering all the patients who are at risk in the doctor's module. This will help doctor to easily identify the high risked patient who need a serious diagnosis. The doctor's module will also have the facility of notifying their patient about their risk via email. Then there is this other notification module which is calculated and generated by the system itself.

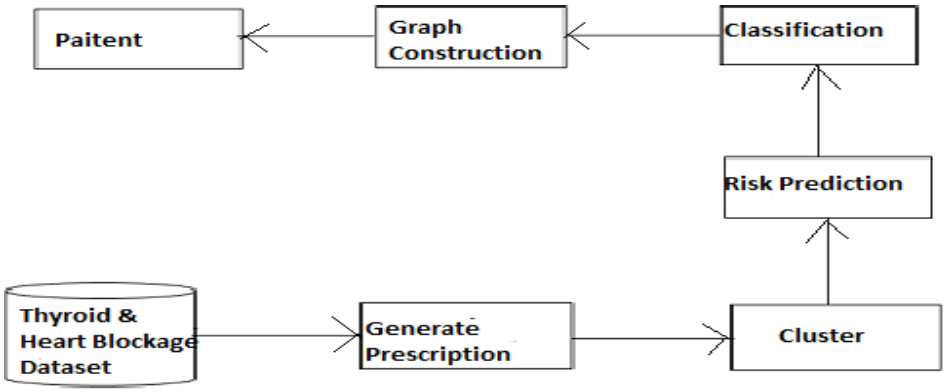

Fig 2:- Architecture of proposed System.

The notification provides the prescription of patient's drugs. Whenever the patient health is at risk, the system will generate a new notification along with its prescription and send it to their profiles.

4.1 Prescription Algorithm

Input: F- Risk prediction n participants

Output: optimized P (Prescriptions) as the computed soft label

Step 1: Initialize Pj

Step 2: Yi : retrieve attribute for F from existing nodes and Di be the death label in record Yi and Dr be the death rate

Step 3: Calculate p for Yi

p = (x * t)n

x = dosage information for p

t = Number of times a day

n = Number of days

Where p is the Prescription details for Yi.

Step 4: Update Pj for j = 1,…..m by:

$Pj(m+1) = \sum_{y}^{1} \sum_{j}^{m} P\,j + p$

4.2 K-Means Clustering Algorithm

Below is the algorithm for k-means clustering:

Let X = {x1,x2,x3,……..,xn} be the set of data points and V = {v1,v2,…….,vc} be the set of centers.

Step 1. Randomly select 'c' cluster centers.

Step 2. Calculate the distance between each data point and cluster centers.

Step 3. Allocate a data point to the cluster center whose distance from the cluster_center is minimum of all the cluster centers.

Step 4. Recalculate the new cluster center using:

$$V_i = (1/cl_i) \sum_{j=1}^{cli} x_i$$

where, 'cli' represents the number of data points in ith cluster.

Step 5. Recalculate the distance between each data point and new obtained cluster centers.

Step 6. If no data point was reassigned then stop, otherwise repeat from step 3.

Let R= {r1,r2,r3……rn}be the set of risk predicted which will act as data points of clusters.

Let 'cl' be the cluster center which is randomly selected.

Step 1: Calculate the distance between data points and cluster center.

Step 2: Assign rn to Cl for all R having minimum distance from Cl.

Step 3: Calculate new data center:

$V_i = (1/Cl_i)\sum_{j=1}^{cli} r_i$

Let M = {m1,m2,m3……mn} be the set of prescription of medicines.

Step 4: Again calculate the distance between the calculated cluster center and data point.If no data point was reassigned then stop,otherwise repeat step 1.

Prescription of each cluster can be calculated as:

$F(v_i) = \lim_{t\to 0} m_{i+1} + t$

5. MATHEMATICAL MODEL

S:- {S0,Sf,I,O,F,SF,DD}

S0:- Initial State of data acceptance from Doctor and Patient

Sf:-Final state of risk.

I:- Data Set D.

D:- X1,X2,Xn.

Xi:- Data item .

O:- Cn no. of graph.

Ci:-Graph C.

n:- No. of Graph.

F:-GetData,Find Risk,GeneratePrescription,CreateGraph

SF:- Getsimilarity

DD:-Group Data D

I:- No of Graph

6. EXPERIMENTAL SETUP AND RESUT

The experiment is carried out by setting two users. Both the user will have the privilege to test the data of undiagnosed patient. This is implemented this system on two datasets:

1. Heart Blockage Dataset: The system trained with data containing 14 attributes having 303 instances. While training the data, proposed system has shown the result mentioned below:

Table 1. Measurements for Heart Blockage Dataset

TEST	RESULT
Correctly Classified Instances	100%
Incorrectly Classified Instances	0%
Kappa statistic	1
Class complexity | order 0	33.1815 bits i.e 0.5925 bits/instance
Mean absolute error	0
Root mean squared error	0
Relative absolute error	0%
Root relative squared error	0%
Total Number of Instances	303

2. Thyroid Dataset: This dataset contains 29 attributes and was trained on 926 instances. After training the instances system will got the following measurements:

Table 2. Measurements for Thyroid Dataset.

TEST	RESULT
Correctly Classified Instances	97.1678 %
Incorrectly Classified Instances	2.8322 %
Kappa statistic	0.9664
Class complexity | order 0	1482.19 bits i.e 3.2292 bits/instance
Mean absolute error	0.0031
Root mean squared error	0.0561
Relative absolute error	3.3246 %
Root relative squared error	25.8392 %
Total Number of Instances	926

The graphical result of performance of both datasets in the proposed system are as follows:

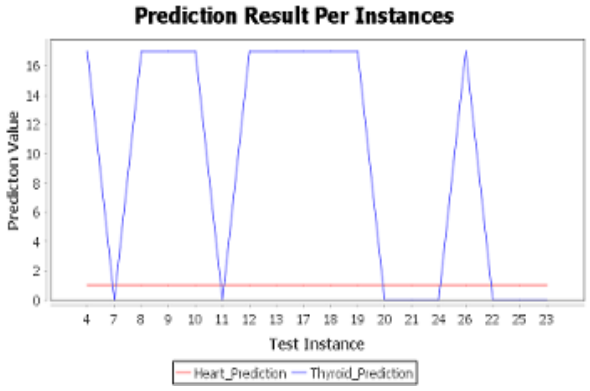

Fig 3:- graphical result of performance of both datasets

This graph will show the risk of each patient to the patient and doctor. Red line showing a risk of heart prediction and blue is showing thyroid prediction.

7. CONCLUSION AND FUTURE WORK

Mining health examination data is challenging due to having personal information provided in the database, it was difficult to find the GHE database where SHG-Health was implemented. But with the help of other datasets, system achieves the goal. System is implemented on heart blockage and thyroid datasets. Hence, Graph based semi supervised learning algorithm is implemented. Along with that proposed system have successfully implemented k-means clustering algorithm to make the cluster of a patient whose condition is worse. And hence, System will notify them and also, provide the prescription.

8. ACKNOWLEDGMENTS

I would like to express my special thanks of gratitude to my guide as well as our HOD & principal who gave me the golden opportunity to do the research on the topic health mining, which also helped me in doing a lot of Research and I came to know about so many new things I am really thankful to them**.**

9. REFERENCES

[1] Ling Chen, Xue Li, Quan Z. Sheng, Wen-Chih Peng, John Bennett, Hsiao-Yun Hu, Nicole Huang, "Mining Health Examination Records—A Graph-Based Approach", *IEEE Transactions on Knowledge & Data Engineering*, vol. 28, no. , pp. 2423-2437, Sept. 2016,

[2] 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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] M. F. Ghalwash, V. Radosavljevic, and Z. Obradovic, "Extraction of interpretable multivariate patterns for early diagnostics," IEEE International Conference on Data Mining, pp. 201–210, 2013

[9] 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.

[10] 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.

[11] 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.797–80