ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

Detecting Risk in Patient based on Health Records- A Graph Approach

Apurva V , B S Umashankar

P.G. Student, Department of Computer Science & Engineering, Don Bosco Institute of Technology, Kumbalagodu, Karnataka, India

Professor, Department of Computer Science & Engineering, Don Bosco Institute of Technology, Kumbalagodu, Karnataka, India

ABSTRACT: In healthcare system, general health check-up is a important part in all the countries. Identifying the patient who is under risk is very challenging which requires precaution and protective involvement. Basically, the unknown data will not determine whether the patient health is good or bad based on the patient report. The data collected about patient may not be true. A graph based algorithm known as semi supervised heterogeneous is proposed. The semi supervised graph algorithm is introduced to reach the goal that will identify the difference between actual data and previously stored data. This algorithm will also helps to predict the risk in a patient. The large amount of health reports will contain remarks, patient ID, laboratory test reports, diagnostics report etc. this health reports which is collected from over the years is known as electronic health report. Finally, this helps to predict whether the health condition of the patient is under risk or not.

KEYWORDS: Semi-supervised learning, labelled data, unlabelled data, heterogeneous graph

I. INTRODUCTION

The Electronic Health Report contains the health information of a single patient. The information contains the identified person, diagnostic reports, medicines and laboratory tests. Health Examination Record (HER) is a kind of Electronic Health Record which contains data of an overall regular health check-up [1]. For an early prevention from death or late diagnosis of a disease, it is necessaryto identify a person who is at risk. To achieve risk calculation which is done by multi class classification problem by means of Cause of Death (COD) used as labels [1]. Main goal of risk calculation successfully is to check if an individual is at risk or not and if yes, predict the related disease of that individual.

The major challenge in the risk prediction is the large amount of unknown data i.e. not all the persons in the records will have a COD label. This record consists of all the cases which vary from healthy to unhealthy conditions. When a healthy person dies without any valid reason mentioned in the record, it acts as an unknown data. Another major constraint of HERs is heterogeneity i.e. it may have an order of unhealthy records and every thing used is mentioned as unknown results. The algorithm makes use of a heterogeneous graph on physical condition test records called Hetero graph, where the test items are of different categories and the links are the temporal relationships between them. The semi-supervised algorithm include all benefits of heterogeneous graph learning [1].

Heterogeneous graph in medical health records means different elements and it will not be having uniform quality. Semi-supervised learning will make use of only unknown data, and supervised learning will make use of known data and large amount of unknown data[1][2]. There are mainly two ways of how the data is stored in HER:

1. General: Only the general information about the patient is stored such as patient name, age, address, phone number, diseases type, blood report etc.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

2. Specified: It depends on for what it is used, may be to transfer test results, or to view diagonis report etc. This data will have certain things which are for particular use only.

The main reason to protect the patient information in HER is that to provide better guidelines and is digitally stored, which is then circulated across the doctors when it is required. The classification methods in healthcare will not consider the problem of unknown data. But some have defined it as low-risk or negative cases, the method which contain unknown data will make use of semi-supervised learning [1].

II. LITERATURE SURVEY

L. Chen, et al. [1] have proposed graph-based classification approach on mining health examination records. It predicts risks for patients considering their health examinations.

M. F. Ghalwash, V. Radosavljevic and Z. Obradovic [2] have proposed an optimization-based method, using which predictive models can be built. They were able to obtain results interpretable by using only optimal number of patterns from the observed data. The shortcoming is that it assumed times series being sampled at regular intervals.

T. P. Nguyen and T. B. Ho [3] have come up with a method based on semi-supervised learning. It integrates many data features for the prediction. It can be applied to study general disease as well as particular disease. The predictions are yet to be validated by experiments.

Ling Chen, et al. [4] have described a data mining-based method for prediction of personal health index based on annual geriatric medical examination records. They expect that the following three aspects will improve the predictions in future:

a) Data semantics: Text fields in the patient records like doctor's suggestions may contain important information and they can be considered for better evaluation effectiveness.
b) Data expansion: Other information sources like smart device recordings can be considered for better results.
c) Dynamic system: Archived database can be made online and dynamic

P. Yang, et al. [5] have discussed the challenges in disease gene prediction. By using an ensemble approach for prediction, their novel method reduces the limitations of individual prediction models.

III. EXISTING SYSTEM

Prediction of risk in patient with the help of only unlabelled data is complicated. Hence, there is no use of predicting risk using unlabelled data.

IV. PROBLEM STATEMENT

The problems of predicting risk using unlabelled data can be prevented by collecting labelled details of the patient and maintaining it in a database. As the records are given to the query of a participant, algorithm calculates whether the patient is at risk or not.

V. PROPOSED SYSTEM

Semi-supervised graph of health is a proposed graph which is based on semi-supervised learning algorithm to analyse the risk at the patients which is having both labelled data and also by using unlabelled data. A resourceful algorithm is designed to identify the risk in patient health. General experiments are carried out on the patient health examination data sets and synthetic dataset.

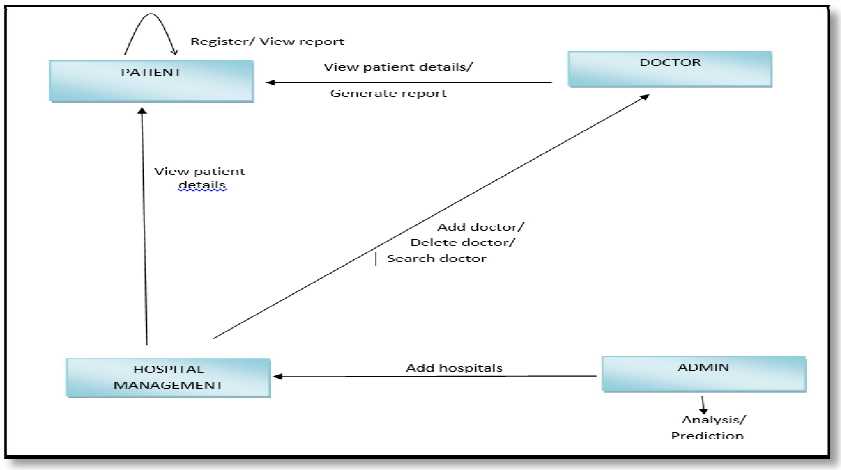
ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

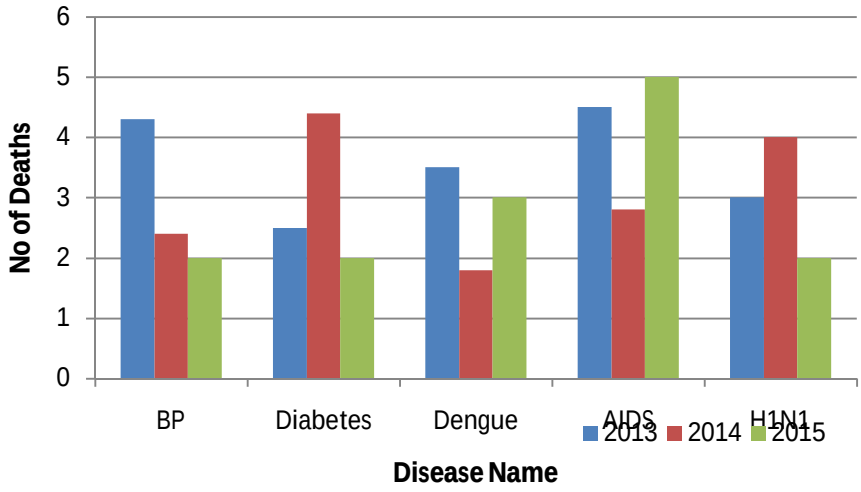


Health system

The architecture of mining health examination is showed above fig, there are mainly four modules in the MHE, namely patient module, doctor module, hospital module and admin module. Here admin module has the full control over the system. Next control of the system is hospital management. In patient module we are having labelled data of particular patient. Doctor module can check the patient and report the information about the patient health to both patient and admin.

VI. EXPERIMENTAL RESULTS

The large amount of health reports will contain remarks, patient ID, laboratory test reports, diagnostics report etc. These health reports which are collected over the years are known as electronic health reports. The health reports helps to predict whether the health condition of the patient is under risk or not. In unlabelled data we are having three years of data. For particular disease, how many people have died is as shown in unlabelled graph. Totally in our database we are having 734 patients' data which is unlabelled i.e. target attribute is unknown.



Unlabelled data graph

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

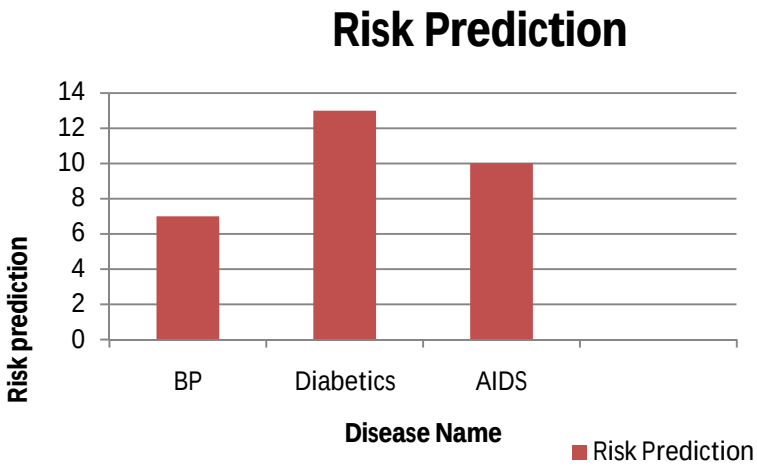
(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

In labelled data, we are having only data where all the details of the patient is known like patient health status, mental health status, treatment given & disease he/she is having etc. by collecting all this information the patient risk is calculated as shown below:
Total number of unlabelled data of patients 734.
Total number of AIDS patients in labelled data is 75.
Total number of sugar patients in labelled data is 96.
Total number of BP patients in labelled data is 52.

In labelled data of patient, he/she is suffering from BP and sugar. So the total number of patient died by BP in labelled data is divided by total number of unlabelled data i.e. 7 percentage of risk the patient is having in BP. Then, total number of patient died by sugar in labelled data is divided by total number of unlabelled data i.e. 13 percentage of risk the patient is having in sugar. Total number of patient died by AIDS in labelled data is divided by total number of unlabelled data i.e. 10 percentage of risk the patient is having in AIDS.



VII. CONCLUSION

In mining of patient health records there are different attributes that need to be counted to calculate the risk in patient's health. These attributes will vary from one patient to another patient. To calculate risk, a semi supervised algorithm which is a graph approach is used. This algorithm will mainly make use of both labelled and unlabelled data for risk prediction. To start with, health reports are represented as a hetero graph which is useful to combine all related disease together. In unlabelled data there will be no attribute value assigned by any doctor and no proper treatment details. In labelled data all the attributes will be checked and then given the treatment details. So we can calculate the risk in patient by using labelled data and also by using unlabelled data. Every time, labelled data is more precise when compared with unlabelled data.

REFERENCES

[1] L. Chen *et al.*, "Mining Health Examination Records—A Graph-Based Approach," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 9, pp. 2423-2437, Sept. 1 2016.
[2] M. F. Ghalwash *et al.*, "Extraction of interpretable multivariate patterns for early diagnostics," in *IEEE International Conference on Data Mining*, pp. 201–210, 2013.
[3] T. P. Nguyen and T. B. Ho, "Detecting disease genes based on semi-supervised learning and protein-protein interaction networks" in *Elsevier - Artificial Intelligence in Medicine* vol. 54, no. 1, pp. 63–71, 2012.
[4] L. Chen, et al., "Personal health indexing based on medical examinations: A data mining approach" in *Elsevier - Decision Support Systems*, vol. 81, pp. 54 - 65, January 2016.
[5] Peng Yang, et al., " Ensemble Positive Unlabeled Learning for Disease Gene Identification" in PLOS ONE, vol. 9, issue , pp. 1 -11, May 2014.