DISEASE DETECTION BASED ON GRAPH AND DATA CLASSIFICATION

S. Rajesh[1], G. Barani[2], V. Suganya[3] K. Seenthamarai[4]
Assistant Professor, Department of Computer Science and Engineering,
Sri Ramakrishna Institute of Technology, Coimbatore, Tamilnadu, India.

Abstract

One of the ultimate goals of life science is to improve our understanding of the processes related to disease. General health examination is an integral part of healthcare in many countries. Identifying the patients at risk is important for early warning and preventive intervention. The fundamental challenge of learning a classification model for risk prediction lies in the collected dataset. Particularly, the collected dataset describes the participants in health examinations whose health related conditions can vary greatly from being healthy to being ill. There is no ground truth for differentiating their states of health. Confidentiality agreements with doctors and patients are also required to obtain patient records. In practice, however, there are many obstacles in the collected dataset because of the limitations of time, cost, and confidentiality conflicts. In this project, We propose an interactive system to predict the patient risks of suffering from certain diseases from the collected dataset by using clustering and classification techniques and thereby giving early warning, and preventive intervention to the patients. This dataset is collected from annual general health check-ups. This helps to save many lives by predicting the disease at an early stage and thereby reducing the cost of medical expenses.

Keywords: Health examination, early warning, Cluster and Classification Techniques, Medical Expenses, HER, Mining, Support Vector Machines.

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, medications, procedures, patient identifying information, and allergies. A special type of EHR is the Health Examination Records (HER) from annual 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 limited 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 as the “highest risk”. The goal of risk prediction is to effectively classify 1) whether a health examination participant is at risk, and if yes, 2) predict what the key associated disease category is. In other words, 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.

II. RELATED WORK

In this section we review existing related studies, namely those on mining health examination data.

A. Data Mining on Health Examination Records

Although Electronic Health Records have attracted increasing research attention in the data mining and machine learning communities in recent years mining general health examination data is an area that has not yet been well-explored, except a few studies on risk prediction such as the chronic disease early warning system. However, none of them considered unlabeled data.

B. Breast cancer survivability prediction using labeled, unlabelled, and pseudo-labeled patient data

Says that [16] survivability of a cancer patient is based on historical patient data. However it is difficult to collect historical patient data because of the confidential agreement between doctors and patients and it is time consuming. To solve this problem we have used semi supervised machine learning algorithms to predict the cancer.

III. DATASETS

A. Real Dataset

The real datasets contain a geriatric health examination dataset and a Cause of Death dataset, linked together via the common attribute Person ID, revealing the association between examination results and main cause of death. Records of participants with non-health-related COD were excluded for the purpose of risk identification.GHE (Geriatric Health Examination) dataset is a de-identified dataset with all private information such as names, contact details, and birth dates removed. The dataset has 230 attributes, containing 262,424 check-ups of 102, 258 participants aged 65 or above, collected during a period of six years (2005-2010). The overall ratio of male to female participants is 1.03:1.

Each de-identified GHE record is represented by a Person ID and the examination results from a wide range of lab tests, physical, the Brief Symptom Rating Scale (BSRS) mental health assessment, the Short Portable Mental Status Questionnaire (SPMSQ) cognitive function assessment, (de-identified) demographics as well as personal health-related habits, such as exercise, eating, drinking, and smoking habits.

For patient demographics, age was first differentiated into five-year bins and Weights (kilogram) and heights (meter) were used to compute the Body Mass Index (BMI) ¼ weight/(height)2. BMI is then recorded as a survivability of a cancer patient with the help of unlabeled data.

B. Mining personal health index from annual geriatric medical examinations

Says that [8] it involves process like data pre-processing, extraction, selection, and model selection. In this they have used personal health index framework that is used to get feedback from the elderly persons and contains cause of death labels and using this we can link patient health index (PHI) to geriatric patient health status and the cause of death.

IV. EXPERIMENTS

Ordinal feature, according to the standard BMI categorization thresholds. For patient habits, most attributes are binary except reason-for-taking-medicine which is a patient reported field. The top seven most frequently reported reasons were extracted as seven binary attributes, each indicating the occurrence of a reason. The rest of the reasons were discarded. For lab tests and physical examinations, there are three fields to record the results, namely observed value, status, and description. The observed values are generally numeric. The status fields indicate whether or not the result of a test is normal. Their values can be either binary or ordinal, depending on the type of tests.

The descriptions are in free text format. We only used the information from the status fields for the following reasons. First, the reference ranges of these items may differ amongst hospitals and the information regarding where an examination was taken is not available in the dataset for privacy reasons. Second, the values for the description fields are mostly missing. The results of five mental health assessment questions were encoded in terms of degree of severity, Normal, mild, medium, and severe. The overall result of the cognitive function assessment (SPMSQ) was scored in terms of the number of questions that were incorrectly answered. It is an ordinal attribute with values “sound” (0-5 scores), “mild” (6-9 scores), "medium" (10-14 scores),and "severe" (above 15 scores) according to the standard categories.

A. Synthetic Dataset

To test the stability of our method, we also generated two groups of synthetic datasets based on the distribution of the processed real datasets. Specifically, we computed the value distributions for a given disease class, a given year, and a given feature for 10 disease classes selected based on COD codes. The cumulative distribution functions (CDFs) were computed, based on which a random number generator was used to select a value for a given disease class, a year, and a feature.

The resulting synthetic datasets are: Balanced datasets: to test the stability with increasing class size in a balanced-class setting, we generated same number of instances for all disease classes and the "unknown" (unlabeled) class with increasing class size in the range and Increasing unlabeled size datasets: to test the stability given increasing scale of unlabeled data, we generated 1,000 instances for every disease class with the "unknown" (unlabeled) class size equals to 1;000.

Four node types were modeled for constructing our Hetero HER graph, namely Record, Physical Test, Mental Assessment, and Profile. Physical Test refers to all the lab tests and physical examination, while Mental Assessment covers both the mental health assessments and cognitive function assessments.

Profile includes all the patient demographics and habits, and Record indicates the artificial nodes created for representing individual records. The numbers of nodes for Physical Test (Test), Mental Assessment (Mental), and Profile types refer to the total number of distinct values of the extracted features for the type data issues, we present a semi-supervised heterogeneity graph based algorithm called SHG-health.

V. METHODOLOGY

A. Data Preprocessing

The dataset contains the huge amount of data. The data may be structured or unstructured in Dataset. If dataset will be unstructured means the preprocessing takes place. In preprocessing phase each and every transaction's are analyzed and determine the parameters are used in the transactions. Thus, the unstructured dataset is converted into structure dataset.

B. TF/IDF (Term Frequency/ Inverse Document Frequency)

Term frequency of term t in document d is defined as the number of times that t occurs in d. Inverse Document Frequency estimates the rarity of a term in the whole document collection. (If a term occurs in all the documents of the collection, its IDF is zero).

C. Weight

Weight is calculated using the product of term frequency and Inverse Document Frequency.

D. Classification

Classification is a general process related to categorization, the process in which ideas and objects are recognized, differentiated, and understood. In the terminology of machine learning, classification is considered an instance of supervised learning where a training set of correctly identified observations is available.

E. Clustering

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar to each other than to those in other groups (clusters). The corresponding unsupervised procedure is known as clustering, and involves grouping data into categories based on some measure of inherent similarity or distance. It is main task of exploring data mining, and a common technique for statistical data analysis, used in many fields.

VI. ALGORITHM FOR COMPARISON

A. Support vector machines

SVM has been adopted as one of our baseline methods. Although SVM with RBF kernel achieved the best results in [8], linear and RBF kernels had very similar performance in our experiments. We only report the results of the linear kernel for its favorable efficiency. The LIBSVM [43] and LIBLINEAR [44] implementations were used in our experiments for the RBF and linear kernels respectively.

B. Nearest neighbor classifier

KNN classifier is a common baseline for graph-based models. K was experimentally set to 1.

VII. CONCLUSION

Mining health examination data is challenging especially due to its heterogeneity, intrinsic noise, and particularly the large volume of unlabeled data. Our proposed graph-based classification approach on mining health examination records has a few significant advantages. First, health examination records are represented as a graph that associates all relevant cases together. This is especially useful for modeling abnormal results that are often sparse. Second, multi-typed relationships of data items can be captured and naturally mapped into a heterogeneous graph. Particularly, the health examination items are represented as different types of nodes on a graph, which enables our method to exploit the underlying heterogeneous sub graph structures of individual classes to achieve higher performance. Third, features can be weighted in their own type through a label propagation process on a heterogeneous graph. These in-class weighted features then contribute to the effective classification in an iterative convergence process. Our work shows a new way of predicting risks for participants based on their annual health examinations. Our future work will focus on the data fusion for the health examination records to be integrated with other types of datasets such as the hospital-based electronic health records and the participants' living conditions (e.g., diets and general exercises). By integrating data from multiple available information sources, more effective prediction may be achieved.

REFERENCES

[1] T. Tran, D. Phung, W. Luo, and S. Venkatesh, "Stabilized sparse ordinal regression for medical risk stratification," Knowl. Inform.Syst., vol. 43, no. 3, pp. 555–582, Mar. 2015.

[2] B. Qian, X. Wang, N. Cao, H. Li, and Y.-G. Jiang, "A relative similarity based method for interactive patient risk prediction," Data Mining Knowl. Discovery, vol. 29, no. 4, pp. 1070–1093, 2015.

[3] Extending association rule summarization techniques to assess risk of diabetes mellitus," IEEE Trans. Knowledge Data Eng., vol. 27, no. 1, pp. 130–141, Jan. 2015.

[4] G. J. Simon, P. J. Caraballo, T. M. Therneau, S. S. Cha, M. R. Castro, and P. W. Li] Mining personal health index from annual geriatric medical examinations," in Proc. IEEE Int. Conf. Data Mining, 2014,pp. 761–766.

[5] Ling Chen, Xue Li,Sen Wang J. Kim and H. Shin, "Breast cancer survivability prediction using labeled, unlabeled, and pseudo-labeled patient data," J. Amer. Med. Inform. Assoc., vol. 20, no. 4, pp. 613–618, 2013.

[6] X. Wang, F. Wang, J. Wang, B. Qian, and J. Hu, "Exploring patient risk groups with incomplete knowledge," in Proc. IEEE Int. Conf.Data Mining, 2013, pp. 1223–1228.

[7] T. P. Nguyen and T. B. Ho, "Detecting disease genes based on semi-supervised learning and protein-protein interaction networks," Artif. Intell. Med., vol. 54, no. 1, pp. 63–71, 2012.

[8] J. M. Wei, S. Q. Wang, and X. J. Yuan, "Ensemble rough hypercuboid approach for classifying cancers," IEEE Trans. Knowl. Data Eng., vol. 22, no. 3, pp. 381–391, Mar. 2010.