IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 3, March 2017

An Improved High Risk Prediction in Health Examination Record Using Data Mining

A. Narmatha[1], R. Abinaya Revathy[2], Mr. S. Radhakrishnan[3]

Bachelor of Technology in Information Technology, Kamaraj College of Engineering and Technology, Virudhunagar[1,2]

Assistant Professor/IT, Kamaraj college of Engineering and Technology, Virudhunagar[3]

Abstract: Medical databases have accumulated huge amounts of information about patients and their medical conditions. Relationships and patterns within the data can provide new medical knowledge. Huge amount of Electronic Health Records (EHRs) are 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, patient identifying information, and allergies. A special type of EHR is the Health Examination Records (HER) from annual general health check-ups. The fundamental challenge of learning a classification model for risk prediction lies in the unlabelled data that constitutes most the collected dataset. Particularly, the unlabelled data describes the participants in health examinations whose health conditions can vary greatly from healthy to very-ill. There is no ground truth for differentiating their states of health. Identifying participants at risk based on their current and past HERs is important for early warning and preventive intervention. Risk means unwanted outcomes such as mortality and morbidity. The proposed system presents a Semi-supervised learning algorithm to handle a challenging multi-class classification problem with substantial unlabelled cases. This algorithm constructs a training set from the diabetes records with unlabelled classes and performs risk analysis with user queries reports. The process shows a new way of predicting risks for participants based on their annual health examinations.

Keywords: Medical database, association rule mining,semi-supervised learning.

I. INTRODUCTION

Nowadays, the data accumulated in medical databases are progressively growing up quickly, this makes extracting hidden knowledge from medical database complex and more time consuming. Analysing these data is critical for medical decision makers and managers. The performance of patient management tasks will be improved by analysing the medical data. Medical data analysis is highly required for the following reasons: 1) Support of specific knowledge-based problem solving activities through the analysis of patient's raw data collected in monitoring, 2) Discovery of new knowledge that can be mined through the analysis of groups of case studies, described by symbolic or numeric descriptors.

Because of these reasons, the usual manual data analysis is not enough especially in case of huge database. The best solution to such problems is using, knowledge discovery in databases (KDD), which has been developed quickly in the last few years. KDD is the process of extracting useful knowledge from large datasets. Data mining is the central step in the KDD process, which deals with the problem of extracting interesting, implicit, and useful relations and patterns in data. The association rule mining is one of the best studied models for pattern discovery in datamining. Extracting knowledge from medical databases can be efficiently done by using association rules. Graph based association rules mining simplifies the process of generating frequent itemsets (symptoms in medical data) by reading the database only once to build an association graph among frequent items (most occurring symptoms).

II. RELATED WORKS

Recently, the knowledge mining applications from medical databases have been increased rapidly. There are two classes of mining techniques applied on medical data: explanatory and exploratory [1]. Explanatory mining refers to techniques that are used for verification or decision making. Exploratory mining is data investigation usually performed at an early stage of data analysis in which an exact mining objective has not yet been set [2]. In the last few years, the number of studies using different techniques of learning on explanatory mining in medical data has been increased progressively. Genetic programming technique has been applied to find out classification rules from medical data sets [3]. Breast cancer survivability has been worked on using AdaBoost algorithms. The fuzzy modelling idea has also been developed on selected features medical data. A system to extract association rules from health examination data has been proposed, after that a case-based reasoning model is used to support the continual disease analysis and management [4]. A different rule mining method with case-based reasoning has been applied recently. Medical data warehouses have been constructed as an extension to the normal medical databases. On the other hand, very few

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

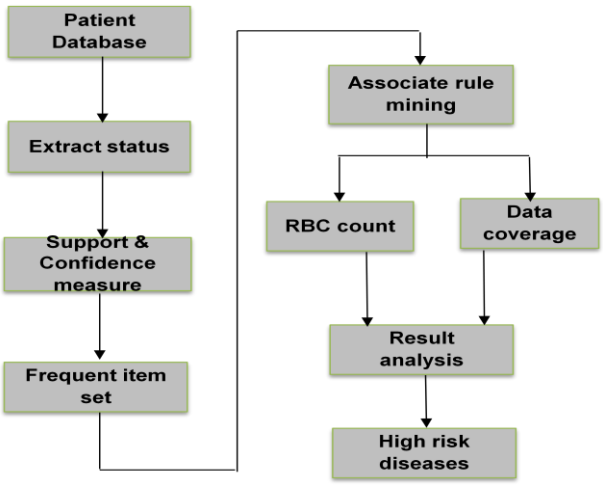
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 3, March 2017

studies talk about the exploratory mining techniques to extract rules from medical databases. And so, this paper deals with exploratory mining technique. Knowledge visualization in the study of hepatitis patients is one of the studies that use exploratory rule mining in the field of medical data [5]. Another study goes to improve visualization by using the functionality of OLAP tools.

III. PROPOSED SYSTEM

The proposed SHG-Health algorithm takes health examination data (GHE) and the linked cause of death (COD) labels as inputs. Its key components are a process of Heterogeneous Health Examination Record (HeteroHER) graph construction and a semi-supervised learning mechanism with label propagation for model training. Given the records of a participant pi as a query, SHG-Health predicts whether pi falls into any of the high-risk disease categories or "unknown" class whose instances do not share the key traits of the known instances belonging to a high-risk disease class. It presents the SHG-Health algorithm to handle a challenging multi-class classification problem with substantial unlabeled cases which may or may not belong to the known classes.

This work pioneers in risk prediction based on health examination records in the presence of large unlabeled data. A novel graph extraction mechanism is introduced for handling heterogeneity found in longitudinal health examination records. By adapting a graph-based approach and exploring the underlying graph structure of health examination records with semi-supervised learning, our method is capable of handling large unlabeled data. To train a disease risk prediction model that can identify high-risk individuals given no ground truth for "healthy" cases, we treated the "unknown" class as a class to be learned from data. To capture the heterogeneity naturally found in health examination items, we constructed a graph called HeteroHER consisting of multi-type nodes based on health examination records. As a preparatory step, all the record values are first discretized and converted into a 0=1 binary representation, which serves as a vector of indicators for the absence/presence of a discretized value. This setting is primarily based on the observation that physicians make clinical judgments generally based on the reported symptoms and observed signs, and secondarily for the reduction of graph density. Every node is typed according to the examination category that its original value belongs to, for example, the Physical tests (A), Mental tests (B), and Profile (C). All the other non-Record type nodes that are linked to the Record type nodes can be seen as the attribute nodes of these Record type nodes. Every attribute (non-Record) type node is linked to a Record type node representing the record that the observation was originally from. The weight of the links is calculated based on the assumption that the newer a record the more important it is in terms. 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. Firstly, 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. Secondly, the values for the description fields are mostly missing.



IV. RESULT

Association rule mining to identify sets of risk factors and the corresponding patient subpopulations that are at significantly increased risk of progressing to diabetes. The system found that the most important differentiator between the algorithms is whether they use a selection criterion to include a rule in the summary based on the expression of the rule or based on the patient subpopulation that the rule covers. Between TopK and BUS, we found that BUS retained slightly more redundancy than TopK, which allowed it to have better patient coverage and better ability to reconstruct the original data base. Also the regular diet information are given according to the patients risk stage which is maintained as an updating diet specification via email.

V. CONCLUSION

Mining health examination data is challenging especially due to its heterogeneity and particularly the large volume of unlabeled data. Despite the lack of direct evidence, early detection through screening is already taking place both by inviting individuals from the general population to come forward for screening and, opportunistically, when individuals perceived to be at high risk of developing diabetes attend for health care (usually primary health care) for other reasons. These activities present opportunities for collecting observational data which, although no substitute for direct RCT evidence, can provide important, circumstantial evidence about efficiency, costs and impact. There is direct evidence that the incidence of diabetes can be reduced in people at high risk of the future development of type 2 diabetes who may

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 3, March 2017

be identified as a result of activities directed towards diabetes detection.

VI. FUTURE ENHANCEMENT

In future work, the classification performance is increased by using SVM classifier. By using this SVM classifier, the high-risk diseases are classified from the patient's extraction results. SVMs are set of related supervised learning methods used for classification and regression. They belong to a family of generalized linear classification. SVM simultaneously minimize the empirical classification error and maximize the geometric margin. SVMs arose from statistical learning theory; the aim being to solve only the problem of interest without solving a more difficult problem as an intermediate step. SVMs are based on the structural risk minimization principle, closely related to regularization theory. It classifies the diseases effectively and provide the patients' health report with high accuracy.

REFERENCES

[1] 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.
[2] T. Tran, D. Phung, W. Luo, and S. Venkatesh, "Stabilized sparse ordinal regression for medical risk stratification,"Knowledge and Information Systems, pp. 1–28, Mar. 2014.
[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, "Predict-ing 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] J. M. Wei, S. Q. Wang, and X. J. Yuan, "Ensemble rough hy-percuboid approach for classifying cancers,"IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 3, pp. 381–391, 2010.
[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 Infor-matics, vol. 51, pp. 8–13, 2014.

DOI10.17148/IJARCCE.2017.63132