An Efficient Classification Approach for Predicting Cause of Death using Mixed Probability Rule Based Algorithm

Pamparaboyena Satyaveni[1], E. Deepthi[2]

Final M.Sc. Student[1], Lecturer[2]

[1, 2] M. Sc Computer Science, Chaitanya Women's PG College, Old Gajuwaka, Visakhapatnam Andhra Pradesh

Abstract:

Now a day's examining the health of each person in the every country is an integral part of healthcare. After examining the health of each person we can identify type of risk to be occurred. The analysis of risk based unlabelled data can be done by using classification approach in the data mining. Particularly we are take unlabelled data contains information related to participants in the health examination whose health condition is vary from great health to very ill. 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 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. In the examination of health we are identifying different states of health without ground truth. So that by predicting risk of each participant by using classification approaches in the data mining. In this paper we proposed Mixed Probability Binary Rule Based Classification Algorithm is used to predict health risk of participate person. By implementing this algorithm we can get efficient classification result and also give better performance.

Keywords: *Classification, Data Mining, Risk Prediction, Cause of Death.*

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. Enormous Amounts of Electronic Health Records (EHRs) composed over the years have provided a rich base for risk examination and forecast. An EHR contains numerically warehoused healthcare info about an individual, such as interpretations, 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 proposal periodic geriatric health examinations as an essential part of their matured care programs. Since clinical care frequently 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 gathered for consistent investigation and defensive purposes, covering a inclusive set of general health measures, all together at a point in time in a methodical way. Identifying contributors at risk based on their current and past HERs is important for early cautionary and preventive intervention. By "risk", we unsolicited outcomes such as mortality and morbidity. In this study we expressed the task of risk forecast as a multi-class classification problem using the Cause of Death (COD) information as labels, concerning the health-related death 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 related 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 conditions that are related to certain exact diseases. Most developed countries have experienced dramatic growth in elderly populations from the beginning of this century. In recent years, this, together with the rising cost of healthcare has created an urgent need for improving predictions and efficient treatment. These programs enable continuous and comprehensive recording of a person's health status, as well as the tracking of his/her health changes.

However, it is always a difficult task for healthcare professionals to provide an overall report on personal health after a comprehensive medical check-up has been performed because of hundreds of the parameters available to be considered. One particular focus of preventive healthcare is risk assessment. The goal is to identify individuals at risk for further investigation or early treatment and intervention. Traditionally, risk assessments have been conducted manually by clinical professionals based on their expertise. These manual assessments have been constrained by the capacity of the human brain to process information within a limited time during the period of an appointment with a patient. Many risk scoring systems have been developed in the field of medicine to assist clinical decision-making. As a general practice in medical research, these methods have been defined based on factors selected with expert knowledge and validated via population based studies. With the advances in computing technology and the availability of EHRs, an increasing number of data mining and machine learning applications have been developed to support healthcare decision making. In recent years methods for clinical risk classification have been developed . However, most existing studies have their focus on EHRs. GHE records and the unique challenges they pose have not yet been well explored. This gap has driven our research to advance risk prediction models for GHE records.

II. RELATED WORK

To the best of our knowledge, the problem of chronic disease hospitalization prediction using machine learning methods is novel. A closely related problem, which has received a lot of attention in the literature, is the re-hospitalization prediction, since around 20% of all hospital admissions occur within 30 days of a previous discharge. Medicare penalizes hospitals that have high rates oreadmissions for some specific conditions that now include patients with heart failure, heart attack, and pneumonia. Examples of work on this problem include . Other related problems considered in the literature are: predicting the onset of diabetes using artificial neural networks developing an intelligent system that predicts, using data-mining techniques, which patients are likely to be diagnosed with heart disease and using data-mining techniques to predict length of stay for cardiac patients (employing decision trees, SVM, and artificial neural networks) or for acute pancreatitis (using artificial neural networks) . We should also mention the Heritage Health Prize, a competition by Kaggle, whose goal was to predict the length of stay for patients who will be admitted to a hospital within the next year, using insurance claims data and data-mining techniques.

Living in modern large cities is impacting our health in many different ways . Primarily due to: (i) stress associated with fast-paced urban life, (ii) a sedentary lifestyle due to work conditions and lack of time, (iii) air pollution, and (iv) a disproportionate number of people living in poverty, urban populations face an increased risk for the development of chronic health conditions . For example, according to the World Health Organization, ambient (outdoor air) pollution was estimated in 2012 to cause 3 million premature deaths worldwide per year; this mortality is due to exposure to small particulate matter of 10 microns or less in diameter (PM10), which cause cardiovascular, respiratory disease, and cancers. In fact, the vast majority (about 72%) of these air pollution-related premature deaths were due to ischemic heart disease and strokes. There is an increasing percentage of the world population facing the adverse health effects of urban living. Specifically, according to the United Nations, 54% of the earth's population resides in urban areas, a percentage which is expected to reach 66% by 2050. It becomes evident that the health of citizens should become an important priority in the emerging smart city agenda. To that end, smart health care –"smart health" as it has been called– involves the use of eHealth and mhealth systems, intelligent and connected medical devices, and the implementation of policies that encourage health, wellness, and wellbeing. It is estimated that by 2020 the smart city market will be worth about $1.5 trillion, with smart health corresponding to 15% of that amount . Additional potential actions smart cities can adopt include ways to improve city life, reduce congestion and air pollution levels, discourage the use of tobacco products and foods high in fat and sugar which increase the risk of chronic diseases, and improve access to health care. Without overlooking the importance of all these population-level measures, our work aims at enabling personalized interventions using an algorithmic data-driven approach.

Through smart health, smart cities and governments aim at improving the quality of life of their citizens. In the state of Massachusetts, the Mass Health program –a combination of Medicaid and the Children's Health Insurance Program– provides health insurance for 1.9 million Massachusetts residents, children in low-income households, low-wage workers, elders in nursing homes, people with disabilities, and others with very low incomes who cannot afford insurance. The state's fiscal year 2018 budget includes approximately $16.6 billion for MassHealth, which is around 37% of the total state budget. Clearly, this is a substantial share of the budget. Consequently, if health care costs can be lowered through smart health, more resources will become available for many other services smart cities can offer. Conversely, if other aspects of smart cities can be improved, the adverse health effects of urban living can be reduced, thus lowering health care costs. This suggests a beneficial feedback loop involving smart health and non-health-related smart city research.

III. PROPOSED SYSTEM

In this paper we are propose mixed probability binary rule based classification algorithm for predicting health risk of participant. To solve the problem of health risk prediction based on health examination of records of participant. Our algorithm takes health examination data and linked cause of death labels as inputs. Its key components process health examination records and predicting disease class. Before processing health examination records we are take the training data set contains information related to test result with type of disease class. By taking those dataset as training data set and predicting examination records of participant. Take the more than one record of participant and processing those records for predicting risk. The predicting health records will consider as testing data set for finding type of disease class. To identify type of disease class we proposed mixed probability binary rule based classification algorithm.

The algorithm combines the advantages of for class discovery and for handling heterogeneity to solve a specific problem induced by evidence-based risk prediction from health examination records. To train a disease risk prediction model that is capable of identifying high-risk individuals given no ground truth for "healthy" cases, we treated the "unknown" class as a class to be learned from data. We incorporated the class discovery mechanism of into our method to handle the "unknown" class. To handle unknown class we propose mixed probability binary rule based classification algorithm for predict type of disease class. The implementation procedure of mixed probability binary rule base classification algorithm is as follows.

Mixed Probability Binary Rule Based Classification Algorithm:

In this module we are implementing mixed probability binary rule based classification algorithm for predicting type of disease classes. By implementing this algorithm we can get best predictive result and also improve the performance. The implementation of steps of mixed probability binary rule base classification algorithm is as follows.

1. Read the training data set contains information related to type of disease class with test results.

2. Read the testing data set contains information related to test result. By taking those testing result we can predict type of disease class.

3. Take the first attribute value from the testing dataset and compare to with training dataset same attribute. If the testing dataset attribute value is greater than equal to training dataset attribute values then put one for that attribute of record. Here we can also consider the testing dataset attribute value is greater than or equal to normal test result we can put one to the particular record.

4. In the comparison process testing data set attribute values less than training dataset attribute value or the testing data set attribute values less than normal test result the put zero as status to particular record.

5. Take the second attribute value from testing data set and perform the step 3, 4 put status of each record with one or zero.

6. This process repeated until the completion of all attributes in training data set and testing dataset values can be converted into in form of zero or one.

7. Take the each attribute value and calculate probability of each attribute related to testing data set attribute.

8. The calculation of probability of each attribute can be done by using following equation.

$Prob_{yes}$= Total number of once/ total number of records.

$Prob_{no}$ =Total number of zeroes /total number of records.

9. Calculate each attribute yes, no probability and find out final yes, no probability of each record. The calculation of final yea, no probability is as follows.

P_{yes} = multiplication of all attributes yes probability.

P_{no} = multiplication of all attributes no probability.

10. After completion of probability calculation we can perform the rule based classification process. The rule based classification process contains If then rules predicting disease class.

11. The rule based classification makes use of a set if then rules for classification. We can express the rule in following form.

If condition then conclusion.

12. In the rule based classification we can take if part of rule is called rule antecedent or precondition.

13. The then part of rule based classification is called rule consequent.

14. The antecedent part of condition consists of one or more attribute tests and these tests or logically And.

15. The antecedent part of our process will take condition as probabilities of yes or no and test dataset values of each attributes and normal test result of each attribute.

16. If all condition of each record satisfies particular disease class of training data set attribute values and take those disease class as predict of risk. Then the predict result is consequent to testing data set result and those participant face the type of disease class.

17. Step 16 will be repeated until total completion records in testing data set.

After completion of this process we can get type of disease that participant will face and also get efficient result. By implementing mixed probability binary rule based classification technique we can get best predict result and also improve performance of system. Because in this algorithm we can calculate probability of each attribute and also generate rule for each record. By performing those two operations we can retrieve more related predict result.

IV. CONCLUSIONS

Mining health examination data is challenging especially due to its heterogeneity, intrinsic noise, and particularly the large volume of unlabelled data. By examining the unlabelled dataset we are using classification technique the data mining. In this paper we are proposed mixed probability binary rule based classification process for predicting type of disease class. In the proposed system we are calculate each attribute probability related to training dataset and using that probability for predicting disease class. In this paper we can also implement the rule based classification approach for identifying type of risk based disease class. In this project we can take two type of dataset for identifying risk. The first data set is training dataset contains information related type of disease class with related test result of attributes. The second data set testing data set is used to identify predict type of disease class using training dataset. By implementing those two processes we can improve efficiency and also get best predict result with the type of disease class.

REFERENCES

[1] M. Woodward. Epidemiology: study design and data analysis. CRC Press, 2013.

[2]. N. Esfandiary, M. R. Babavalian, A.-M. E. Moghadam, and V. K. Tabar. Knowledge discovery in medicine: Current issue and future trend. Expert Systems with Applications, 41(9), Jul. 2014.

[3]. J.-Y. Yeh, T.-H. Wu, and C.-W. Tsao. Using data mining techniques to predict hospitalization of haemodialysis patients. Decision Support Systems, 50(2):439– 448, Jan. 2011.

[4]. F.Wang, P. Zhang, B. Qian, X.Wang, and I. Davidson. Clinical risk prediction with multilinear sparse logistic regression. In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 145–154, New York, USA, 2014. ACM.

[5]. T. Tran, D. Phung, W. Luo, and S. Venkatesh. Stabilized sparse ordinal regression for medical risk stratification. Knowledge and Information Systems, 43(3):555–582, mar 2015.

[6]. J. Wiens, E. Horvitz, and J. V. Guttag. Patient Risk Stratification for Hospital-Associated C. diff as a Time-Series Classification Task. In Neural Information Processing Systems, pages 476–484, 2012.

[7]. Y. Mao, W. Chen, Y. Chen, C. Lu, S. Louis, M. Kollef, and T. C. Bailey. An integrated data mining approach to real-time clinical monitoring and deterioration warning. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 1140–1148, Beijing, China, 2012. ACM

[8]. H. Neuvirth, M. Ozery-Flato, J. Hu, J. Laserson, M. S. Kohn, S. Ebadollahi, and M. Rosen-Zvi. Toward personalized care management of patients at risk: the diabetes case study. In SIGKDD, pages 395–403, California, USA, 2011. ACM.

[9]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.

[10] 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.

[11] J. M. Wei, S. Q. Wang, and X. J. Yuan, "Ensemble rough hypercuboid approach for classifying cancers," IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 3, pp. 381–391, 2010.

[12] 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.

[13].Q. Nguyen, H. Valizadegan, and M. Hauskrecht, "Learning classification models with soft-label information." Journal of the American Medical Informatics Association : JAMIA, vol. 21, no. 3, pp. 501–8, 2014.

[14] G. J. Simon, P. J. Caraballo, T. M. Therneau, S. S. Cha, M. R. Castro, and P. W. Li, "Extending Association Rule Summarization Techniques to Assess Risk of Diabetes Mellitus," IEEE Transactions Knowledge and Data Engineering, vol. 27, no. 1, pp. 130–141, 2015.

BIOGRAPHIES:

Pamparaboyena Satyaveni is student in M.Sc. (Computer Science) in Chaitanya Women's pg college, Old Gajuwaka, Visakhapatnam. She has received his Degree B. Sc (mscs) from MVR Degree College, chinagantyada, Visakhapatnam. Her interesting areas are data mining, network security and cloud computing

E. Deepthi is lecturer in Department of M.Sc. Computer Science in Chaitanya Women's PG College. Old Gajuwaka, Visakhapatnam, Andhra Pradesh. She Receive her Ph.D. from ph.d pursuing from KL university .Her research areas include Network Security and Computer Networks. She published 10 international journals and she was attended number of conferences and workshops.