ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Special Issue 19, April 2017

SURVEY ON RISK STRATIFICATION ON PUL HEALTH CARE DATA USING ESHG ALGORITHM

V.Premalatha[#1], G.Kalpana*[2].

[#]Dept. of Computer Science and Engineering SRM UNIVERSITY, Ramapuram, Chennai, Tamilnadu, India.
[1]premalatha1308@gmail.com
* Dept. of Computer Science and Engineering, SRM UNIVERSITY, Ramapuram Chennai, Tamilnadu, India,
[2]g.gkalpana@gmail.com

Abstract— **Data mining in health analytics is a challenge for understanding a classification type for predicting the risk. Risk prediction lies in the data which is in the form of unlabeled data type and it gives the majority of data which are collected from the health analytical techniques. . The collected data sets which don't have COD (Cause Of Death) factor is named as unlabeled data Here the unlabeled data defines the participants physical fitness and it significantly changes over a time. For this complication there is no static rule for describing the participant physical fitness. Therefore an effective iterative method of ESHG(Enhanced Semi supervised Heterogeneous Graph) is defined to identify the progressive development of the unlabeled data. Health analysis from the different data source using information technology of data mining would provide the effective and efficient technique of risk prediction which helps the participants to know the risk factors and the risk situation that are associated to some explicit disease. The main objective is to meeting the production targets of risk prediction of the individual participants.**

KEY WORDS: Risk prediction, Unlabeled data, Cause of Death, Semi supervised learning algorithm.

I. INTRODUCTION

Risk prediction models are used in health examination and analytical method of decision making and are used to help the participants to make an informed choice about the treatment. The main objective in predicting the risk is to determine whether levels of risk is low, medium or high. In their analysis the risk factors are purely based on the participants that are thought to be associated with the health events of interest. The major challenge is the large amount of unlabeled data and it is treated as negative data. Many of the health examination records stating that 92.6 % of the participants dataset do not have the COD label. so that it would become a multiclass learning method of unlabeled type of data. Classification of data is to predict the target class for each case in the data to predict the risk levels. And it provides the supervised function to learn the best prediction. Combination of relevant data or information from two or more data sources that provide more accurate description than any of the individual data sources. Semi supervised learning algorithm provides the combination of heterogeneous learning of graph and classification of data shows the considerable performance gain in the collected health examination data and it is used to handle the heterogeneous data in longitudinal manner with substantial negative data called unlabeled data. ESHG algorithm uses both positive called labeled data and unlabeled data to classify the risk at different levels.

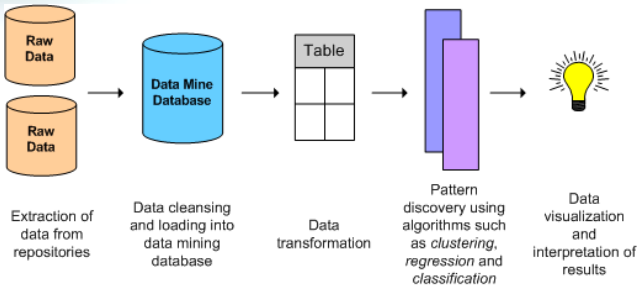


Fig.1 Data Mining in Information Technology

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Special Issue 19, April 2017

II. PROBLEM STATEMENT AND SURVEY

A. *Problem Definition*

Data mining in health examination is competitive and challenging nowadays, especially because of its heterogeneity and huge amount of unlabeled data. In existing system, all focused only labeled data, and they treat unlabeled data simply as negative case. Earlier model predicts the risk of the participants based on the annual health records. It is not focusing on the data fusion with other type of datasets such as electronic health examination records with respect to time analysis. Previous handling of risk prediction might not be tuned for efficiency. Both this paper will reveal the desirable efficiency and effectiveness in both time and stability of collected datasets.

B. *Related Work*

[1] a large volume of breast cancer patient data is required to build the predictive models. In the machine learning or data mining domain, the types of data are categorized as 'labeled' and 'unlabeled'(features without labels). For patient data related to breast cancer survivability, the labels tags a patient as 'survived' if they survived for a specific period or 'not survived' if they did not. Accumulating a large volume of labeled data is time consuming, costly and it requires confidentiality agreements

[2] describes the risk stratification, which aims to stratify the patient data into a homogeneous groups according to some risk evaluation criteria. And it an important task in modern medical informatics. Good risk evaluation is the key to good personalized care plan design and delivery. The typical procedure for risk stratification to first identify the set of risk- relevant medical features (also named as risk factors), and then construct a predictive model to estimate the risk scores for individual patients. However, due to the heterogeneity of 'patients' clinical conditions, the risk factors and the importance can vary across different patient groups. Therefore a better approach is to first segment the patient cohort into a set of homogeneous groups with consistent clinical conditions, namely the risk groups, and then develop group - specific risk prediction model.

[3] discussed about a system, the effective incentive scheme is proposed to stimulate the forwarding cooperation of nodes in VANETs. In a coalitional game model, every relevant node cooperates in forwarding messages as required by the routing protocol. This scheme is extended with constrained storage space. A lightweight approach is also proposed to stimulate the cooperation.

[4] The use of telehealth technologies to remotely monitor the patients suffering chronic disease may enabled preemptive treatment of worsening health conditions before a significant deterioration in the subject's health status occurs, requiring hospital admission. The objective of this study has to develop and validate a classification algorithm for the early identification of patients', with the background of chronic obstructive pulmonary disease (COPD), who appears to be high risk of an imminent exacerbation event. The algorithm attempts to predict the patient's condition one day in advance, based on a comparison of their current physiological measurements against the distribution of their measurements over the previous month. The proposed algorithm, which uses a classification and regression tree (CART) has been validated using telehealth measurement data recorded from patients with moderate/severe COPD living at home. The CART algorithm can classify home telehealth measurement data in to either a 'low risk' or 'high risk' category with 72% accuracy, 85% specificity, 63%sensitivity. The algorithm was able to detect a 'high risk' condition one day prior to patients actually being observers as having a worsening in their COPD condition, as defined by symptom and medication records. This study highlights the potential usefulness of automated analysis of telehealth data in the early detection of exacerbation events among COPD patients.

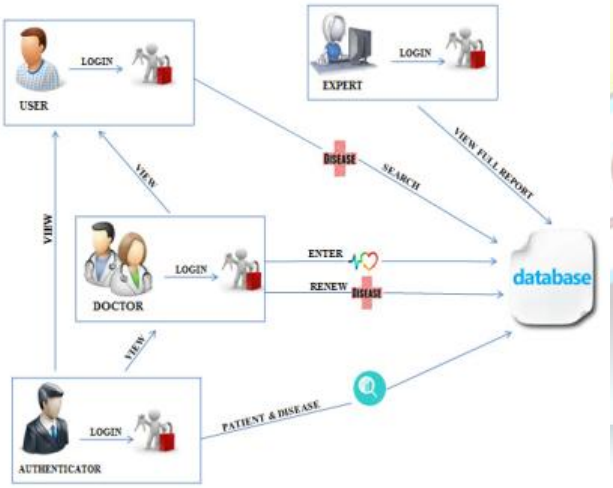
[5] Early detection of patients with elevated risk of developing diabetes mellitus is critical to the improved prevention and overall clinical management of these patients. This paper aim to apply association rule mining to electronic medical records(EMR) to discover sets of risk factors and their corresponding sub populations that represent patients at particularly high risk of developing diabetes. Given the high dimensionality of EMRs, association rule mining generates a very large set of rules which needs to summarize for easy identification. The reviewed association rule set summarization techniques and conducted a comparatives evaluation to provide guidance regarding applicability, strength and

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Special Issue 19, April 2017

weakness. The propose extension to incorporate risk of diabetes into the process of finding an optional summary. It summarized the described sub population at high risk of diabetes with each method having its clear strength. For this purpose, extension to the Bottom-Up Summarization (BUS) algorithm produced the most suitable summary. The sub population identified by this summary covered most high-risk patients, had low over lap and were at very high risk of diabetes.

III. ANALYSIS OF FRAMEWORK

Data mining and machine learning techniques, with the goal of knowledge discovery and deriving data driven insights from various data sources, has played a more and more important role in medical informatics. Effective data mining approaches have been applied in many medical problems including personalized medicine, disease modeling, cohort study, comparative effectiveness researches, etc.



In this study, here the patients and doctors can able to login to the system to view the clear details of the individual with controlled access to the centralized server. Data authentication is given by the admin of the system. Data experts do their analysis and retrieve the reports as per the essential requirements of the health examination.

1) Algorithm

The k-anonymity algorithm is used to context of data table releases . This paper reiterate their definition and then proceed to analyze the merits and shortcomings of k-anonymity as a privacy model. The k-anonymity model has three entities: individual, whose privacy needs to be protected ; the database owner , who controls a table in which each row (also named as records or tuple) describes exactly one individuals; and the attacker. The k-anonymity model makes two major assumptions.

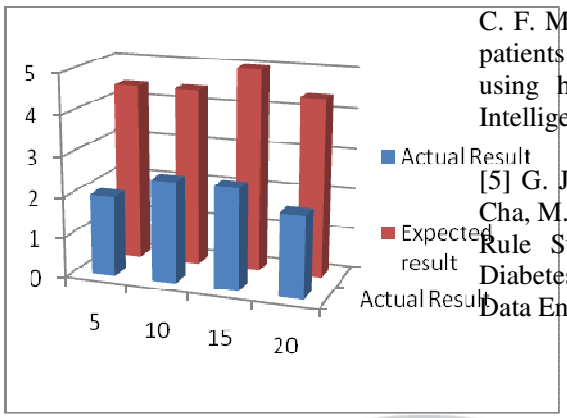
- The database owner is able to separate the columns of the table into a set of quasi-identifiers, which are attributes that may appear in external tables the database owner does not control, and a set of private columns, the values of which need to be protected. We prefer to term these two sets as public attributes and private attributes, respectively.

- The attacker has full knowledge of the public attribute values of individuals, and no knowledge of their private data. The attacker only performs linking attacks. A linking attack is executed by taking external tables containing the identities of individuals, and some or all of the public attributes. When the public attributes of an individual match the public attributes that appear in a row of a table released by the database owner, then we say that the individual is linked to that row. Specifically the individual is linked to the private attribute values that appear in that row. A linking attack will succeed if the attacker is able to match the identity of an individual against the value of a private attribute.

IV. EXPECTED OUTCOME

This shows the expected outcome of risk prediction in health examination record . And it is consider as a suitable method of risk prediction with high efficiency.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Special Issue 19, April 2017



V. CONCLUSION

In this survey we identified the better prediction model using unlabeled data with high efficiency and effectiveness in health care analytical techniques. However, some are some obstacles when collecting patients confidentiality conflicts. Therefore further research should be taken care on the electronic health care examination using digital data like finger print, biometrics test to ensure the detailed description of the individuals even in the case of emergency condition.

REFERENCES

[1] J. Kim and H. Shin, "Breast cancer survivability prediction using labeled, unlabeled, and pseudo-labeled patient data," Journal of theAmerican Medical Informatics Association : JAMIA, vol. 20, no. 4, pp.613–618, 2013.

[2] X. Wang, F. Wang, J. Wang, B. Qian, and J. Hu, "Exploring patientrisk groups with incomplete knowledge," IEEE InternationalConference on Data Mining, pp. 1223–1228, 2013.

[3] Christo Ananth, M.Muthamil Jothi, A.Nancy, V.Manjula, R.Muthu Veni, S.Kavya, "Efficient message forwarding in MANETs", International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE), Volume 1,Issue 1, August 2015,pp:6-9

[4] 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.

[5] 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.