Automatic Risk Prediction in Health Examination Records Using Mining

Mrs.P.Saraswathi, AP
Department of Information Technology,
Velammal College of Engineering and Technology
Madurai - Tamil Nadu, India

Ms.T.Aarthy, Ms.K.Sri Kavi Bharathi and Ms.S.Ezhilarasi
Department of Information Technology,
Velammal College of Engineering and Technology
Madurai - Tamil Nadu, India

Abstract **- General health examination is an integral part of healthcare. Identifying the participants at risk is important for early warning and preventive intervention. The fundamental challenge of classification model for risk prediction is the unlabeled data that constitutes the majority of the collected dataset. The unlabeled data describes the participants in the health examination record whose health conditions can vary greatly from healthy state to very-ill. There is no ground truth for differentiating their states of health. Huge amounts of Electronic Health Records (EHRs) collected over the years have provided a rich base for risk analysis and prediction. The proposed system presents an improved classification known as C4.5 algorithm to handle a challenging multi-class classification problem with substantial unlabeled cases. C4.5 constructs a decision tree starting from a training set with the divide and conquers strategy. C4.5 is based on the information gain ratio that is evaluated by entropy. The information gain ratio measure is used to select the test features at each node in the tree. Such a measure is referred to as a feature (attribute) selection measure. The attribute with the highest information gain ratio is chosen as the test feature for the current node. The results show that the proposed systems achieve effectiveness and efficiency on both real health examination datasets and synthetic datasets.**

Keywords— Labeled and Unlabeled data; Classification; Prediction; C4.5 Algorithm.

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,laboratorytests,diagnosticreports,medications,procedures,patient identifying information, and allergies.A special type of EHR is the Health Examination Records (HER) from annual general health check-ups.HERs are collected for regular surveillance and preventive purposes,covering a comprehensive set of general health measures all collected at a point in timein a systematic way.

Identifying participants at risk based on their current and past HERs is important for early warning and preventive intervention. The goal of risk prediction is to effectively classify

- whether a health examination participant is at risk, and if yes,
- predict what the key associated disease category is.

A good risk prediction model should be able to exclude the low-risk situations thereby clearly identifing the high-risk situations that are related to some specific diseases.A fundamental challenge is the large quantity of unlabeled data. Most existing classification methods on healthcare data do not consider the issue of unlabeled data. They either have expert-defined low-risk or control classes or simply treat non-positive cases as negative. Methods that consider unlabeled data are generally based on Semi-Supervised Learning (SSL) that learns from both labeled and unlabeled data. Amongst these SSL methods, only handle large and genuinely unlabeled health data.However, unlike our scenario, both methods are designedfor binary classification and have predefined negative cases . A closely related approach is Positive and Unlabeled (PU)learning, which can be seen as a special case ofSSL with only positive labels available. other key challenge of HERs is heterogeneity. It demonstrates the health examination records of Participants with test items in different categories. This example shows that

- A patient may have a sequence of irregularly time-stamped longitudinal Health records, each of which is likely to be sparse in terms of abnormal results.
- Test items are naturally in categories, each conveying different semanticsand possibly contributing differently in risk identification.

II. EXISTING SYSTEM

The Existing SHG-Health algorithm takes health examination data (GHE) and the linked cause of death (COD) 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 complex multi-class classification problem with substantial unlabeled cases which may or may not belong to the known classes. This work succeeds in making risk predictions 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 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. 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.

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.

Disadvantages:

- Multivariable prediction model is over fitting.
- High risk prediction occurs.
- Only handle large unlabeled data.

III. PROPOSED SYSTEM

The proposed system presents a new classification approach in Health examination records by using C4.5 algorithm. This algorithm constructs a decision tree starting from a training set in which decision tree is a tree data structure consisting of decision nodes and leaves. The Decision tree is one of the classification techniques which are done by the splitting criteria. The decision tree is a tree structure that classifies instances by sorting them based on its features value. Each node in a decision tree represents a feature in an instance to be classified. The instances are classified from beginning based on its feature value. Decision tree generates the rule for the classification of the data set. The value of a split point depends on how well the separation of classes takes place. Numerous splitting indices have been proposed in the precedent to evaluate the quality of the split. The C4.5 is the statistic Classifier. C4.5 algorithm uses gain ratio for selection of feature and to construct the decision tree. It handles both continuous and discrete features. C4.5 algorithm is widely used because of its quick classification and high precision. It is needed to perform supervised data mining on the target data set. This narrowed down the choice of classifiers to only few, classifiers can handle numeric data as well as give a classification. Hence selecting C4.5 decision tree learning became obvious. The attribute evaluation was also performed in order to find out the gain ratio and ranking of each attribute in the decision tree learning. In case for some data set data mining could not produce any desirable result then finding the correlation coefficient was resorted to investigate if relation between attributes. Due to dealing with huge dataset, a collection of decision tree classification algorithm has been decided. The advantages of C4.5 algorithm is significantly, so it can be choose. But its efficiency must be improved to meet the dramatic increase in the demand for large amount of data. With the set of given records, each record has the same structure, consisting of a number of attribute/value pairs. It determines the decision tree on the basis of answers to questions about the non-category attributes predicts correctly the value of the category attribute. The current attribute node which has the maximum value of information gain which has been generated, and the root node of the decision tree is obtained in this way. Having studied carefully, each node in the selection step of test attributes there are logarithmic calculations available, and in each time these calculations have been performed previously too. The efficiency of decision tree generation can be impacted when the dataset is large.

Advantages:

- It is used to guide treatment towards a more personalized.
- It is useful for modeling abnormal results that are often sparse.
- Predicting risks for participants.

IV. METHODOLOGY

Software:

- Language : Java
- IDE : Net Beans
- Data Base : MySql

Algorithm:

- Association Rule Mining
- C4.5 algorithm

Association rule mining:

All rules that correlate the presence of one set of items with that of another set of items.

Support: Support denotes the frequency of the rule within transactions. A high value means that the rule involve a great part of database.

Confidence: Confidence represents the percentage of transactions involving A which contain also B. It is an estimation of conditioned probability.

C4.5:

C4.5 is a well-known algorithm used to generate a decision trees. The decision trees generated by the C4.5 algorithm can be used for classification, and for this reason, C4.5 is also referred to as a statistical classifier.

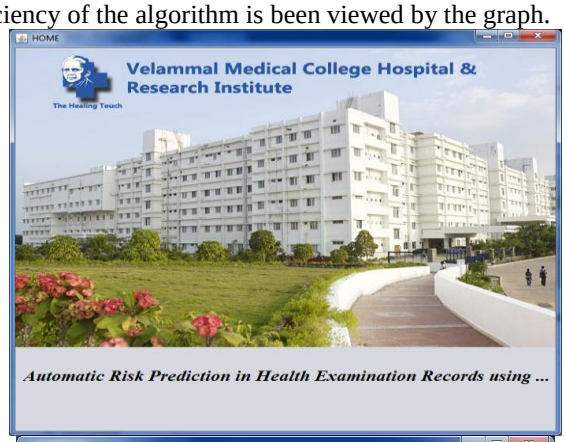
- Handling training dataset with missing values of attribute set.
- Handling differing cost attributes
- Pruning the decision tree after its creation

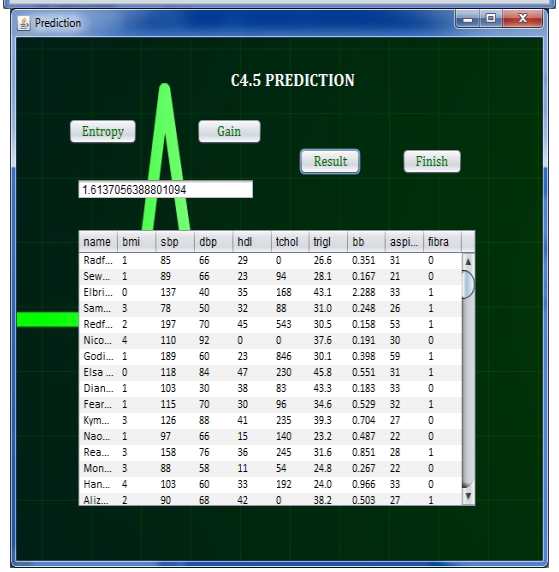
V.MODULES

1. Admin
2. User
3. Health system

1. ADMIN:

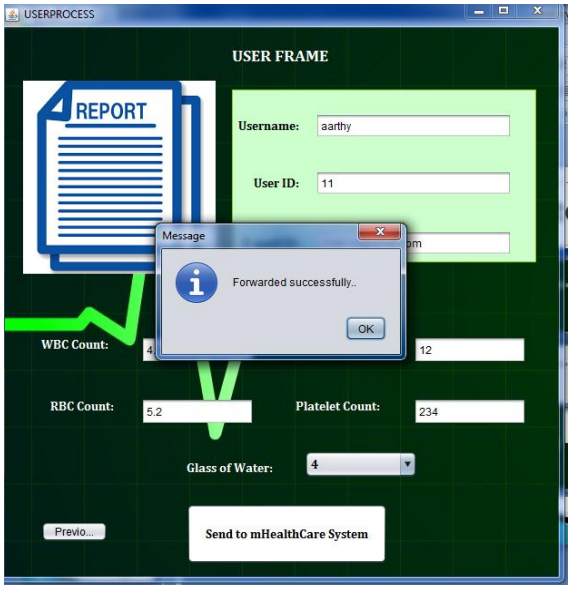
The admin module enables the doctor to view all the medical records previously stored and allows him/her to perform mining process. Association rule mining associates the data set using the threshold values and rule mining does the process of determining the High risk patients in the total record set. The two parameters used are Support and Confidence. The Support determines the union of number of patients visited and the number of times they have visited the hospital. The high risk patients are predicted in this phase which enables efficient point of care to the patient who are at risk. The C4.5 algorithm which checks the efficiency of the result generated by the Association rule mining algorithm using two parameters namely Entropy and Gain. Finally the efficiency of the algorithm is been viewed by the graph.





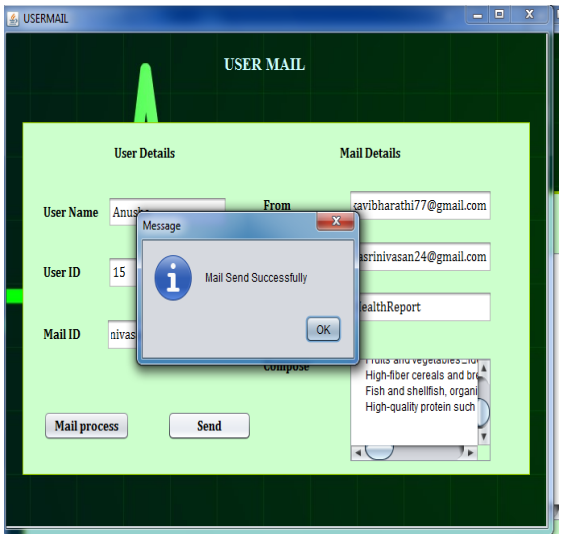
2. USER:

The user (patient) can login in to the portal using the username, password and captcha and if they are the new user then they can register as a new user and can login in to the portal. In the portal username, Id, mailId will be displayed along with the parameters such as WBC Count, RBC Count, Platelet Count, Hemoglobin Count and glass of water. The user can give all the values for the parameters and by clicking finish the parameter values are been stored in the database. The user can also able to view the previous report of him/her by which they can able to understand about their health status.



3. HEALTH SYSTEM:

The Health system allows the admin (ie., Doctor) to view all the reports that are been entered through the User module. The records will be displayed with the Id, username with the parameters. It generate the result as whether the patient is normal or abnormal. Based on the generated result mail will be sent to the user mailId. The appointment time and date can also be added to the mail content.



VI.CONCLUSION

The system proposed a new classification approach to predict the high risk rate using C4.5 algorithm.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 whether they use a selection criterion for including a rule condition in the summary based on the expression of the rule or based on the patient subpopulation that the rule covers.This algorithm constructs a decision tree starting from a training set in which decision tree is a tree data structure consisting of decision nodes and leaves. Entropy Computation is used to create

compact decision trees with successful classification. The size of the decision tree, the performance of the classifier is based on the entropy calculation. So the most precise entropy can be applied to the particular classification problem.The different entropies based approach can be applied in any classification problem.The results show a new way of predicting risks for participants based on their annual health examinations.

VII. FUTURE WORK:

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 a set of supervised learning methods used for the purpose of 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 is to solve only the problem of interest without being solving a complex problem as an intermediate stage. SVMs which are based on the basic structural risk minimization principle, closely related to regularization theory. It classifies the diseases effectively and provides the patients' health report with high accuracy.

VIII.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.

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

[7] 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.

[8] L. Chen, X. Li, S. Wang, H.-Y. Hu, N. Huang, Q. Z. Sheng, and M. Sharaf, "Mining Personal Health Index from Annual Geriatric Medical Examinations," in2014 IEEE International Conference on Data Mining, 2014, pp. 761–766.

[9] S. Pan, J. Wu, and X. Zhu, "CogBoost: Boosting for Fast Cost-sensitive Graph Classification,"IEEE Transactions on Knowledge and Data Engineering, vol. 6, no. 1, pp. 1–1, 2015.

[10] M. Eichelberg, T. Aden, J. Riesmeier, A. Dogac, and G. B. Laleci, "A survey and analysis of Electronic Healthcare Record standards," ACM Computing Surveys, vol. 37, no. 4, pp. 277–315, 2005.

[11] C. Y. Wu, Y. C. Chou, N. Huang, Y. J. Chou, H. Y. Hu, and C. P. Li, "Cognitive impairment assessed at annual geriatric health examinations predicts mortality among the elderly," Preventive Medicine, vol. 67, pp. 28–34, 2014.

[12] "Health assessment for people aged 75 years and older," http://www.health.gov.au/internet/main/publishing.nsf/ Content/mbsprimarycarembsitem75andolder, accessed: 2015-05-03.

[13] "Health checks for the over-65s," http://www.nhs.uk/Livewell/ Screening/Pages/Checksover65s.aspx, accessed: 2015-05-03.

[14] L. Krogsbøll, K. Jørgensen, C. Grønhøj Larsen, and P. Gøtzsche, "General health checks in adults for reducing morbidity and mortality from disease (Review),"Cochrane Database of Systematic Reviews, no. 10, 2012

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

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

[17] H. Huang, J. Li, and J. Liu, "Gene expression data classification based on improved semi-supervised local Fisher discriminant analysis,"Expert Systems with Applications, vol. 39, no. 3, pp. 2314– 2320, 2012.

[18] T. P. Nguyen and T. B. Ho, "Detecting disease genes based on semi-supervised learning and protein-protein interaction networks," Artificial Intelligence in Medicine, vol. 54, no. 1, pp. 63–71, 2012.

[19] V. Garla, C. Taylor, and C. Brandt, "Semi-supervised clinical text classification with Laplacian SVMs: An application to cancer case management,"Journal of Biomedical Informatics, vol. 46, no. 5, pp. 869–875, 2013.

[20] X. Wang, F. Wang, J. Wang, B. Qian, and J. Hu, "Exploring pa-tient risk groups with incomplete knowledge," IEEE International Conference on Data Mining, pp. 1223–1228, 2013.