ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering
(IJARMATE) Vol. 3, Special Issue 2, April 2017

AN IMPROVED HIGH RISK PREDICTION IN HEALTH EXAMINATION RECORDS USING MINING

T.VENNILA
.M.E. COMPUTER SCIENCE AND ENGINEERING
INDRA GANESAN COLLEGE OF ENGG
Trichy, TAMILNADU
yelroenila@gmail.com

D.INDRA DEVI/ ASSO. PROF.
COMPUTER SCIENCE AND ENGINEERING
INDRA GANESAN COLLEGE OF ENGG
TRICHY, TAMILNADU
csehodigce@gmail.com

***Abstract*——** General health examination is an integral part of healthcare in many countries. Identifying the participants at risk is important for early warning and preventive intervention. The fundamental challenge of learning a classification model for risk prediction lies in the unlabeled data that constitutes the majority of the collected dataset. Particularly, the unlabeled 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. In this paper, we propose a graph-based, semi-supervised learning algorithm called SHG-Health (Semi-supervised Heterogeneous Graph on Health) for risk predictions to classify a progressively developing situation with the majority of the data unlabeled. An efficient iterative algorithm is designed and the proof of convergence is given. Extensive experiments based on both real health examination datasets and synthetic datasets are performed to show the effectiveness and efficiency of our method.

Keywords—component; formatting style styling; insert (key words)

I. INTRODUCTION

UGEamounts of Electronic Health Records (EHRs)collected over the years have provided a rich base forrisk analysis and prediction. An EHR contains digitally stored healthcare informa-tion about an individual, such as observations, laboratorytests, diagnostic reports, medications, procedures, patientidentifying information, and allergies. A special type ofEHR is the Health Examination Records (HER) from annualgeneral health check-ups. For example, governments suchas Australia, U.K., and Taiwan, offer periodicgeriatric health examinations as an integral part of theiraged care programs. Since clinical care often has a specificproblem in mind, at a point in time, only a limited andoften small set of measures considered necessary are col-lected and stored in a person's EHR. By contrast, HERs arecollected for regular surveillance and preventive purposes,covering a comprehensive set of general health measures.

Identifying participants at riskbased on their current and past HERs is important for earlywarning and preventive intervention. In this study we formulated the task of risk prediction asa multi-class classification problem using the Cause of Death(COD) information as labels, regarding the health-relateddeath as the "highest risk". The goal of risk prediction is toeffectively classify 1) whether a health examination partici-pant is at risk, and if yes, 2) predict what the key associateddisease category is. In other words, a

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering
(IJARMATE) Vol. 3, Special Issue 2, April 2017

good risk predictionmodel should be able to exclude low-risk situations andclearly identify the high-risk situations that are related tosome specific diseases.A fundamental challenge is the large quantity of unla-beled data. For example,92:6%of the 102,258 participants inour geriatric health examination dataset do not have a CODlabel. The semantics of such "alive" cases can vary fromgenerally healthy to seriously ill, or anywhere in between.In other words, there is no ground truth available for the"healthy" cases. If

we simply treat this set of alive cases asthe negative class, it would be a highly noisy majority class.On the other hand, if we take this large alive set asgenuinelyunlabeled, as opposed to cases with known labels removed,it would become a multi-class learning problem with largeunlabeled data.Most existing classification methods on healthcare datado not consider the issue of unlabeled data. They either haveexpert-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 labeledand 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.

While the unlabeled set Uin a PU learning problem is similar to our alive set,its existing applications in healthcare only address binaryclassification problem. Nguyen et al. introduced a multi-class extension called mPUL; however, their methodused a combined set of negative and unlabeled example,while in our case negative example is not available.The other key challenge of HERs isheterogeneity. It demonstrates the health examination records of Participantp1in three non-consecutive years with test items in differ-ent categories (e.g., physical tests, mental tests, etc.) andabnormal results marked black. This example shows that1) a participant may have a sequence of irregularly time-stamped longitudinal records, each of which is likely to besparse in terms of abnormal results, and 2) test items arenaturally in categories, each conveying different semanticsand possibly contributing differently in risk identification.Therefore this heterogeneity should be respected in themodeling.

II. EXISTING SYSTEM

The Existing SHG-Health algorithm takes health examination data (GHE) and the linked cause of death (COD) labels described in Section 5.1 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 present 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.

Disadvantages: Multivariable prediction model is over fitting.High –risk prediction occurs. Only handle large unlabeled data.

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering
(IJARMATE) Vol. 3, Special Issue 2, April 2017

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 is done by the splitting criteria. The decision tree is a flow chart like a tree structure that classifies instances by sorting them based on the feature (attribute) value. Each node in a decision tree represents a feature in an instance to be classified. All branches denote an outcome of the test, each leaf node hold the class label. The The decision tree is a flow chart like a tree structure that classifies instances by sorting them based on the feature (attribute) value. Each node in a decision tree represents a feature in an instance to be classified. All branches denote an outcome of the test, each leaf node hold the class label. The instances are classified from starting based on their feature value. Decision tree generates the rule for the classification of the data set. The Existing SHG-Health algorithm takes health examination data (GHE) and the linked cause of death (COD) labels described in Section 5.1 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 present 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.

IV. CLASIFICATION

Step 1. Binarization: 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 indi-cators for the absence/presence of a discretized value. Specifically, real values, such as age, are first binned into fixed intervals (e.g., 5 years). Then, all the ordinal and categorical values are converted into binary representa-tions.

Step 2. Node Insertion: Every element in the binary repre-sentation obtained in Step 1 with a value "1" is modeled as a node in our HeteroHER graph, except that only the abnormal results are modeled for examination items (both physical and mental). This setting is primarily based on the observation that physicians make clinical judgements generally based on the reported symptoms and observed signs, and secondarily for the reduction of graph density.

Step 3. Node Typing: 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) in Fig. 1. In addition, a new type of nodes is introduced to represent individual records such as r_{11}, r_{12}, and r_{13} in the same figure. 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. In other words, categories A, B, and C in Fig. 1 can be regarded as the attributes of the Record type at a schema level. This leads to a graph schema with a star shape as shown on the right of Fig. 3 below, which is known as a star schema [34]. Note that types can often be hierarchically structured and thus choosing the granularity of node type may require domain knowledge or be done experimentally.

Step 4. Link Insertion: Every attribute (non-Record) type node is linked to a Record type

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering
(IJARMATE) Vol. 3, Special Issue 2, April 2017

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 term.

V.EXPERIMENT RESULT

Table 5.1 Evaluation on Disease class prediction

	Precision	Recall/Sensitivity	F-Score
Ours	96.24 1.60	**43.93** 1.11	**60.32** 1.23
Ours-Chi2	**99.33** 0.36	43.02 1.40	60.02 1.41
Ours-Gaus	96.99 1.32	43.69 1.14	60.23 0.86
SVM	89.00 10.19	0.49 0.36	0.98 0.71
KNN	37.52 1.48	25.62 1.30	30.45 1.36
GNetMine	0.00 0.00	0.00 0.00	0.00 0.00
GGSSL	5.21 0.02	100.00 0.00	9.91 0.03

Table 5.2 Evaluation of Binary Predicition

	Macro-Precision	Macro-Recall
Ours	89.14 0.56	89.62 0.38
Ours-Chi2	**90.58** 0.19	**90.73** 0.15
Ours-Gaus	89.55 0.56	90.30 0.41
KNN	21.12 1.49	59.92 2.50
SVM	52.50 39.41	63.33 30.55
GNetMine	-	-
GGSSL	0.11 0	9.09 0

Table 5.3 Extracted Hetro HER Graph statistics

	Data set	# people	# nodes					# links to Record nodes				Density
			Record	Test	Mental	Profile	Total	Test	Mental	Profile	Total	
Real	GHE@10class	26,771	73,642	55	26	55	73,778	601,062	119,952	523,387	1,244,401	0.1242
Synthetic	(100,100)	1,100	3,013	55	26	55	3,149	28,071	4,611	22,552	55,234	0.1348
	(300,300)	3,300	9,054				9,190	84,052	13,982	68,053	166,087	0.1349
	(500,500)	5,500	15,092				15,228	139,736	23,134	113,231	276,101	0.1345
	(1000,1000)	11,000	30,201				30,337	280,412	46,655	227,932	554,999	0.1351
	(1000,3000)	13,000	35,463				35,599	323,101	55,263	265,067	643,431	0.1334
	(1000,5000)	15,000	40,695				40,831	365,005	63,974	301,661	730,640	0.1320
	(1000,10000)	20,000	53,674				53,810	469,575	85,167	393,486	948,228	0.1299
	(1000,15000)	25,000	66,841				66,977	575,360	106,694	485,914	1,167,968	0.1285
	(1000,20000)	30,000	79,979				80,115	682,022	128,357	579,269	1,389,648	0.1278

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 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. 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 lassification problem. The results show a new way of predicting risks for participants based on their annual health examinations

VI. CONCLUSION

The system proposed a new classification approach to predict the high risk rate

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

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.

.

