RFMiner: Risk Factors Discovery and Mining for Preventive Cardiovascular Health

Yao Xiao, Ruogu Fang
School of Computing and Information Sciences, Florida International University, Miami, Florida
yxiao009@fiu.edu, rfang@cs.fiu.edu

***Abstract*—Cardiovascular disease is one of the leading causes of death in the United States. It is critical to identify the risk factors associated with cardiovascular diseases and to alert individuals before they experience a heart attack. In this paper we propose *RFMiner*, a risk factor discovery and mining framework for identifying significant risk factors using integrated measures. We provide the blueprints for accurately predicting the possibility of heart attacks in the near future while identifying notable risk factors - especially the factors which are not well recognized.**

I. Introduction

Cardiovascular disease is one of the leading causes of death in the United States. It is critical to identify risk factors for cardiovascular diseases such as heart attack and provide alerts to individuals at risk well in advance. Previous research has singled out health conditions like obesity, diabetes, and smoking as important risk factors for heart attacks [1]. However, with the advent of big medical data and connected health, large-scale datasets with broad coverage in both time and space become available, which provide an opportunity to discover risk factors that were not possible to identify in small-scale datasets. In this paper, we propose *RFMiner*, a risk factor discovery and mining framework for identifying significant risk factors using integrated measures. The experimental results demonstrate that this approach can be performed to identify cardiovascular diseases such as heart attacks.

Specifically, in this framework, we aim to accurately predict the possibility of heart attacks in the near future while identifying notable risk factors, especially the factors which are not well recognized. The contributions of our work include: (1) A cascaded classifier to improve the precision and recall for the unbalanced dataset which outperforms the state-of-the-art results; (2) Discovery of novel risk factors by integrating various interestingness measures.

II. Preventive Cardiovascular Health

A. Dataset

The dataset for this paper comes from the Behavioral Risk Factor Surveillance System (BRFSS) [2], the largest continuously conducted health survey system in the world. For each year, there are about 400,000 records in the BRFSS open dataset and we apply the data from the year 2014. With respect to heart attacks, our goal is to unveil new risk factors to assist diagnose and prevent heart attacks.

B. Data Preprocessing and Attribute Filtering

Based on the BRFSS codebook, we select "CVDINFR4" as our class attribute to identify whether a survey respondent is "Ever Diagnosed with Heart Attack?". To handle the missing data, we remove the instances for lost records or refused answers and the attributes with high missing rates with a threshold of 30%. To reduce the computational complexity, we also remove apparently irrelevant attributes and the sub-attributes of the main circumstances. After preprocessing and down-sampling, there are 77 out of 265 attributes and about 20,000 instances remaining.

C. Heart Attack Prediction using Baseline Classifiers

We first experiment with several simple classifiers - the result of the best eight classification cases with 10-fold cross validation can be seen in Table I. Naive Bayes (NB) demonstrates the best recall at 0.723, while the precision is low at 0.183, which indicates false positives are too high. In contrast, Random Forest (RF) achieves the highest precision of 0.712, with a recall of 0.481, and an AUC of 0.908. Although, the overall performance of RF is favorable compared to other classifiers, the low recall is not preferred for disease screening purposes in a real-world setting.

TABLE I
Classification results of 8 base classifiers, with the best performances in bold font.

Type	Recall	FPR	Precision	F1	AUC	Accuracy
KNN	0.076	0.009	0.347	0.124	0.697	92.50%
NB	**0.723**	0.209	0.183	0.292	0.825	78.68%
J48	0.347	0.015	0.593	0.438	0.712	94.57%
Logistic	0.120	0.005	0.611	0.201	0.838	94.18%
AdaBoost	0.493	0.014	0.690	**0.575**	0.885	95.57%
Bagging	0.370	0.016	0.599	0.458	0.877	94.66%
RF	0.481	0.013	**0.712**	0.574	**0.908**	**95.65%**

D. Cascaded Classifiers

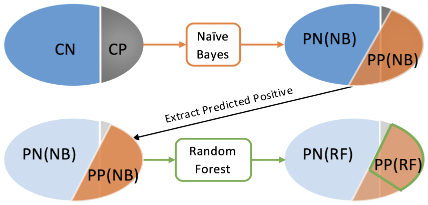

Fig. 1. Illustration of the cascaded classifier combining Naive Bayes (NB) and Random Forest (RF). Condition positive (CP) and condition negative (CN) mean the actual label for the data. PP indicates predicted as positive, PN indicates predicted as negative.

To address the imbalanced data in the testing stage and to build a classification model that can boost the performance in terms of both precision and recall, we design a cascaded classifier that can combine the advantages of both Naive Bayes and Random Forest. First, we use Naive Bayes with the pre-processed data, which can maintain the majority of the instances positive for heart attack. Next, we apply Random Forest to the extracted subset to exclude the false positives. Figure 1 illustrates the idea of the cascaded classifier for imbalanced data classification. The higher the PP becomes the better performance the classifier preforms.

Table II shows the prediction performance using the cascaded classifier in comparison to the individual classifiers. We also compare our results with the state-of-the-art method [3] using an integration of logistic regression for feature selection and an artificial neural network for prediction by reimplementing their method and applying it to our dataset. Although the recall decreases to some extent, the overall F1 score improves over both individual classifiers with decent recall and precision.

TABLE II
PERFORMANCE OF THE CASCADED CLASSIFIER COMPARED TO THE INDIVIDUAL CLASSIFIERS. BEST PERFORMANCES ARE IN BOLD FONT.

	Recall	Precision	F1
LR+ANN [3]	0.177	0.606	0.270
Naive Bayes	**0.723**	0.183	0.292
Random Forests	0.481	0.712	0.574
Cascaded	0.53	**0.726**	**0.613**

III. NOVEL RISK FACTORS DISCOVERY

A. Interestingness Measure and Rank Lists Integration

We use three interestingness measures to explore new risk factors: Odds Ratio (OR), Lift, and Kulczynsky. Since each interestingness measure reflects a unique ranking criterion, integrating these ranked lists would optimize the top-ranked risk factors that are most crucial to trigger the disease. To identify the most important risk factors from all potential factors, we first narrow down the number of risk factors in each list for prediction performance evaluation. We maintain the top $M = 35$ features from each ranked lists to retain a decent number of risk factors for pruning.

The next step is to combine these three lists into one to get the integrated ranked list. We assign each of the three ranked lists l_i, $(i = 1, 2, 3)$ with a list-weight α_i, where i indicates the ranked list index. The list-weight $\alpha_i = \frac{1}{2} \cdot \ln \frac{F_1^i}{1 - F_1^i}$ is a logarithmic function that based on the calculated $F1$-score from the classification result. We also assign each feature j in the ranked lists a weight $w_j^i = \frac{m + 1 - r_j^i}{m + 1}$ based on the rank r_j^i of this feature in the ranked list i. For example, w_3^2 means the weight of the third feature in the second ranked list. The final score $score_j = \sum_{i=1}^{3} \alpha_i \cdot w_j^i$ of each feature j is calculated by the list-weight and the individual feature weight. Based on the final score of the features, the integrated ranked list is generated, which can provide a comprehensive representation of heart attack risk factors.

Table III shows the performance of three interestingness measures and the integrated ranked lists using only the top 35 attributes. It shows that by integrating the different ranked lists established from distinct interestingness measures, complimentary criteria to select the top risk factors can be fused to yield a more representative ranked list of important risk factors. We find a performance improvement in terms of the F1 score of 5% compared to the best performance of the individual list.

TABLE III
THE PERFORMANCE OF CASCADED CLASSIFIER USING TOP 35 ATTRIBUTES FROM THE RANKED LISTS ESTABLISHED USING OR, LIFT, KULCZYNSKI, AND THE INTEGRATED RANK LIST. THE BEST PERFORMANCES ARE IN BOLD FONT.

	Recall	Precision	F1
OR	0.517	0.657	0.579
Lift	0.531	0.668	0.529
Kul	**0.571**	0.598	0.585
Integrated	0.569	**0.673**	**0.617**

B. Novel Risk Factor Discovery

From the integrated rank list, we confirm a number of established risk factors and discover several novel risk factors that have not yet been reported in the literature. Heart attack risk factors include stress (MENTHLTH) [4]; cardiovascular diseases such as coronary heart disease (CVDRHD4) and stroke (CVDSTRK3); cancer (CHCOCNCR), skin cancer (CHCSCNCR); and pulmonary disease (CHCCOPD1) and kidney disease (CHCKIDNY). Furthermore, factors like USENOW3 (tobacco use), DIABETE3 (diabetes), WEIGHT2, and any age-related factors are all well-established [1], [5]. In addition to that, we discover a number of novel risk factors, which include employment status (EMPLOY1), educational level (EDUCA), seat belt usage (SEATBELT), teeth removal (RMVTEETH3), veteran status (VETERAN3), and blindness (BLIND), which would suggest important social-economical determinants for the health conditions.

IV. CONCLUSION

In this paper, we have proposed a framework *RFMinder* for preventive cardiovascular health using a large-scale behavioral risk factor dataset with the use case of heart attack prediction. We boost the prediction performance using a cascaded classifier which mingles the individual classifiers with high precision or recall. We also propose a method for novel risk factors discovery by the integration of ranked lists using various interestingness measures. Extensive experiments on a large-scale dataset demonstrate a significant improvement when using the cascaded classifier for disease prediction.

REFERENCES

[1] B. Dahlöf, “Cardiovascular disease risk factors: Epidemiology and risk assessment,” *The American Journal of Cardiology*, 2010.

[2] C. for Disease Control and P. (CDC), “Behavioral risk factor surveillance system survey data,” Atlanta, Georgia, 2014.

[3] A. Wang and Others, “A logistic regression and artificial neural network based approach for chronic disease prediction: A case study of hypertension,” *IEEE International Conference on Internet of Things*, 2014.

[4] A. Neylon, C. Canniffe, S. Anand, C. Kreatsoulas, G. J. Blake, D. Sugrue, and C. McGorrian, “A global perspective on psychosocial risk factors for cardiovascular disease,” *Progress in Cardiovascular Diseases*, 2013.

[5] D. Buchan, N. Thomas, and J. Baker, “Novel risk factors of cardiovascular disease and their associations between obesity, physical activity and physical fitness,” *Journal of Public Health Research*, vol. 1, no. 1, 2012.