RESEARCH ARTICLE | JANUARY 10 2019

A data mining approach to predict risk of cardiovascular FREE

Shaopeng Ma; Xiong Chen

AIP Conf. Proc. 2058, 020014 (2019)
https://doi.org/10.1063/1.5085527

A Data Mining Approach to Predict Risk of Cardiovascular

Shaopeng Ma[1, a)], Xiong Chen[1, b)]

[1]*Department of electronic engineering, Fudan University, Shanghai 200433, China.*

[a)]spma13@fudan.edu.cn
[b)]Corresponding author email: chenxiong@fudan.edu.cn

Abstract. Cardiovascular disease is now increasingly threatening to humanity. The accurate prediction of patients' condition is significant to early prevention. This paper describes our research about how to predict patients' risk of cardiovascular disease by processing their physical examination reports. We use five items (systolic pressure, diastolic pressure, triglyceride, high-density lipoprotein cholesterol and low-density lipoprotein cholesterol) to quantizer this risk in our research. To extract useful information from the medical records, we use natural language processing (NLP) method. To conserve the sentence into digital data, we use term frequency-inverse document frequency (TF-IDF) algorithm to extract major information from medical reports. Principal component analysis (PCA) algorithm is used to reduce the high dimension of text information data. Additionally, we extracted easy-transform numerical features and category features. Combining all these features, we use the xgboost algorithm to make final predictions. The results turn out to be well that the mean square error and relative error can be restricted to an acceptable low level.

Key words: Medical data mining, XGBoost, NLP.

INTRODUCTION

Cardiovascular disease is one of the main causes of death in most countries. Millions of patients suffer great pain from it and have to pay much for expensive medical treatment. Early screening and timely medical intervention can always play an efficient role to preventing conversion of these chronic diseases to severe form. Cardiovascular risk prediction is a problem to predict a patient's probability of having cardiovascular by processing a wide range of relative data. Most patients could have escaped if they can get accurate rick prediction of cardiovascular disease before worsening of their conditions. The key issue in the field of cardiovascular disease prevention is to give an accurate prediction whether a person is probable to have this disease. Moreover, this research is also significant in the field of medical document management if we can fill the missing data efficiently and precisely.

Empirical methods to estimate risk of cardiovascular disease are now much mature 1 2. Doctors examine, read medical records, and make predictions by the knowledge they absorbed during medical school. Nevertheless, obviously this approach is the most expensive and least efficient for patients.

Subsequent researchers also use medical sensors to collect patients' chemical or electric signals, and then they can detect the abnormal feature and deduce patients' health condition through these data 3. These methods are always proposed or summarized by medical researchers, which can guarantee high interpretability and always play an important role to find out the pathogenesis of disease. Five among all these signals are found to be effective predictors of cardiovascular risk, which are systolic pressure, diastolic pressure, triglyceride, low-density lipoprotein cholesterol (LDL) cholesterol and high-density lipoprotein cholesterol (HDL) cholesterol concentrations456. The weakness of this method is the difficulty to use the specific medical devices.

Benefiting from the magnificent development of computer science and microelectronics, nowadays researchers have capacity to deal with much higher dimensional and much bigger data. A new interdisciplinary subject of medical data mining has emerged. Much interest of processing medical signals using data mining technology has been expressed in recent years 7. Obviously, this method is the cheapest and the most efficient way to get the accurate

International Conference on Frontiers of Biological Sciences and Engineering (FBSE 2018)
AIP Conf. Proc. 2058, 020014-1–020014-7; https://doi.org/10.1063/1.5085527
Published by AIP Publishing. 978-0-7354-1786-1/$30.00

report. The current major problem is how to get higher accuracy and interpretability. Artificial neural network has been applied to identify people at high risk of dyslipidemia 8, which can get pretty good results. However, they used much personal information, such as education level, marital status, lifestyle behaviors and so on. This information is out the field of narrow medical research and is difficult to get, so it is not an efficient way to be used in actual medical application.

We propose a method to make predictions of patients' systolic pressure, diastolic pressure, triglyceride, LDL and HDL concentrations, which can be predictors of cardiovascular disease. We use the recently proposed data-mining algorithm XGBoost to solve this problem, which is more efficient and cheaper. In addition, we only use patient's medical records data, which is more limited in the medical fields and easier to collect.

METHODS

Natural Language Processing

Natural language processing (NLP) is now a very hot and challenging research field. Some mature systems have been established to understand and process language information. The general procedure starts from the tokenization, which means splitting a sentence into a sequence of lexemes. It is especially necessary for Chinese language processing, because Chinese words in a sentence are not separated from each other by space. Next step is the part-of-speech tagging (POS tagging), which means assign the proper part-of-speech to each word given the context information. The frequently used models are hidden Markov model, Maximum Entropy model and Conditional Random Fields model.

Then we need to recognize the entity phrase in a sentence, which is especially important in our research. Intuitively the first thing for a doctor to understand a medical record is extracting the named entity, such as organs diseases and symptoms. Jason P.C. Chiu and Eric Nichols propose a novel Bidirectional LSTM-CNNs architecture model to solve the named entity recognition (NER) task and achieve a pretty good result 9. The package spacy we used to realize the NER task is based on this architecture.

Another necessary part is dependency parsing (DP) analysis, which can deduce the syntactic structure of a sentence. In other words, we can get the interaction of different parts of one sentence and it is a key step to understand natural language for computers. Finally, we can transform a natural sentence into a more constructive form, so it is easier to be quantified.

Multiple Additive Regression Tree

Decision tree algorithm works in a series of if-then procedures, which can be visualized as a tree structure. Multiple Additive Regression Tree (MART, sometimes also named as Gradient Boosting Decision Tree) is a popular boosting regression algorithm, which can be seen an ensemble of classification and regression trees (CART). It is a widely used data-mining algorithm in diverse research fields, such as traffic prediction 10, meteorological research 11, and financial research 12.

$$\hat{y}_i = \sum_{k=0}^{K} f_k(x_i), \tag{1}$$

Where K is the number of trees, f is the decision function of the ith tree, and $\hat{y}_i$ is the estimation result.

We fix what we have learned and add a new tree, which can fit the negative gradient of loss function at each step.

$$r_{ti} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=\sum_{k=0}^{t-1} f_k(x_i)}, \tag{2}$$

Where r_{ti} is the target of the ith example when we build the tth tree, L is the loss function, and y_i is the fitting target 13.

If considering the second order information, we can rewrite the loss function.

$$\sum_i \left[l\left(y_i, \hat{y}_i^{t-1}\right) + \nabla f^t(x_i) + \frac{1}{2}\nabla^2 f^t(x_i)^2 \right], \tag{3}$$

Where l is the loss function of true value y_i and the corresponding prediction $\hat{y}$ at the tth step? The XGBoost 14 15 algorithm use this function as optimization goal, which can reach higher accuracy, higher speed and more robustness. Another advantage is that it can process both continuous and discrete values. We use xgboost algorithm to regress the expected value by information extracted from the raw data.

EXPERIMENT

Dataset Introduction

We consider this problem on the dataset collected by a physical examination institution, which can be downloaded from Tianchi. This dataset has 38199 entries, and each is a patient's physical examination report. All these reports consist of 2806 kinds of items, including patient ID, five prediction targets (systolic pressure, diastolic pressure, triglyceride, high-density lipoprotein cholesterol and low-density lipoprotein cholesterol) and 2800 diverse physical examination items' results. All these items except prediction targets and patient ID have undergone data masking, so we don't know each item's particular meaning, only to have a set of words sequences and some undefined figures.

Data Cleaning

First of all, we need to drop out items that contain obvious mistakes which are out bound of acceptable range. Then we need to find out numerical features hidden in character strings. As most of items are stored as string in the raw data and some of them are polluted by meaningless characters or various encoding style, we use regular expression to match number-like object out of raw string.

Additionally, we need to select features that seem to be universal, because there are 2002 physics examination items less than 156 patients taking part in. The fourth thing we do is to choose some category-like features and transfer them from string to category. For instance, some reports are 'No abnormity', 'No obvious abnormity' or 'Normal', and some reports are 'Negative' or the symbol '-'. We represent these values that mean the same but appear not as one same specific numbered category manually.

NLP

We choose out items that have text features to do natural language processing. Two popular software packages spacy and Stanford Corenlp are used in our research. But the existing pre-trained universal open source language model on the Internet is usually based on news or blog corpus and limited to recognize persons, organizations, locations, dates and so on. So, we have to extract the major and frequent medical terms out from our data manually and build a dictionary including organs symptoms and some frequently used descriptions. Then we can build a mini corpus to train the named entity recognition model.

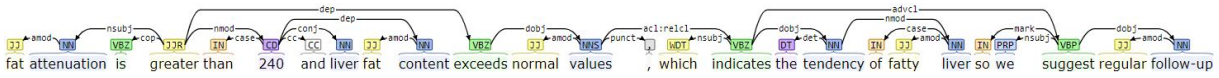

FIGURE 1. Stanford NLP Dependency Parsing

Figure 1 is one example result of dependency parsing. The line with arrow describes the relationship between two words. This sentence is quite long, and it can be translated as fat attenuation is greater than 240 and liver fat content exceeds normal values (the higher fat attenuation value is, the higher fat liver has), which indicates the tendency of fatty liver so we suggest regular follow-up. It is not just a description of a symptom, but added with some suggestions and explanations, which is common in medical records. We can extract the major information from these long sentences by a group of words corresponding to the interested entities. Finally, we can process the text feature as a category text feature.

We also use another algorithm to transform the text items to numerical features. TF-IDF algorithm is a method to vectorize sentence based on the frequency that word emerges. The first step is to create a vocabulary dictionary of words that appear not too much time, which means it will ignore high-frequency words such as 'normal' and keep the relative low-frequency words such as 'hepatic adipose infiltration'. So, it can play the role of a filter to reserve the most important word in sentences. It is especially useful in medical record data processing, because few parts of patients have abnormal symptoms in most cases. Then we can transform each sentence into a vector based on whether it has these key words in the dictionary, which is somewhat like the one-hot encoding. Finally, we reduce the dimensionality of vectors using principal component analysis (PCA) algorithm. This low-dimensional vector we finally get can reserve main information of this item.

XGBoost

After all the preprocessing above, we now have 359 numerical features, 178 text features, 189 category features. Not all these features above are decoupled, some of them treated as text and category features at the same time. The 178 features are processed a combination of dimension reduced matrix.

The smaller items are the more relative error their prediction results are probable to generate. To reduce the sensitivity of large numbers, log function is applied to the five items we want to regress. Because log function is a monotone increasing concave function, small number can be mapped more precisely as big number less precisely.

$$x' = \log(1 + x) \tag{4}$$

We feed x' to the model and use the mean square error (MSE) of x' to evaluate the prediction results. We use the software package XGBoost to realize our method.

RESULTS AND DISCUSSION

MSE of Different Feature Combinations

We choose different combinations of features to feed the model and evaluate the MSE of different items on test dataset.

TABLE 1. The mse of five items by different combination of features

	systolic pressure	diastolic pressure	triglyceride	HDL cholesterol	LDL cholesterol	mean
text	0.01386	0.01786	0.09660	0.01281	0.03852	0.03593
numerical	0.01494	0.01831	0.07592	0.01094	0.03157	0.03033
category	0.01836	0.02149	0.11716	0.01439	0.04202	0.04268
text+numerical	0.01287	0.01676	0.07337	0.01024	0.03022	0.02869
text+category	0.01379	0.01780	0.09558	0.01275	0.03844	0.03567
numerical+category	0.01371	0.01745	0.07422	0.01045	0.03051	0.02926
all	0.01288	0.01662	0.07297	0.01027	0.03010	0.02857

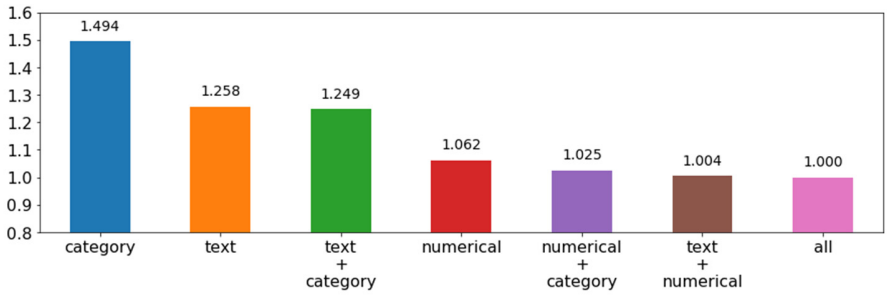

FIGURE 2. The relative mean mse of five items feeding different combination of features

The result, given in Table 1, is the MSE of five items by different combination of features. The MSE of systolic pressure, diastolic pressure and HDL cholesterol is below 0.02 in almost all cases while LDL cholesterol varies between 0.030 and 0.042. The MSE of triglyceride ranges from 0.073 to 0.117.

We treat the result all features combined as the benchmark to evaluate the experiment results of other cases. The relative MSE of different combinations is shown in Figure 2. We find that if only one type of features used, the numerical features work best, then the text features, and the category features work worst. Their mean MSE are 1.494, 1.258 and 1.062 times as much high as the benchmark respectively. It seems that the categorizing preprocessing of text features cannot extract much useful information to make predictions. The numerical features contribute most to make an accurate prediction.

When we use two features to feed the model, the combination of text and category features can get an MSE of 0.0357, the combination of numerical and category can reach 0.0293, and the combination of text and numerical can reach 0.0286. It seems that the adding of category features does not make the results much better, but it is still useful. The combination of text and numerical features can get a promotion about 6% lower than only numerical features used. When using all three features, we can get the best result of 0.0286.

Effect of Log Function

TABLE 2. Relative error of five items in two different conditions

	systolic pressure		diastolic pressure		triglyceride		HDL cholesterol		LDL cholesterol	
	log	no log	log	no log	log	no log	log	no log	log	no log
LQ[a]	0.0356	0.0368	0.0413	0.0409	0.1404	0.1422	0.0507	0.0519	0.0692	0.0663
Medians	0.0775	0.0766	0.0855	0.0855	0.2846	0.3083	0.1085	0.1086	0.1413	0.1425
Mean	0.0913	0.0921	0.1041	0.1054	0.3853	0.4420	0.1370	0.1393	0.2040	0.2092
UQ[b]	0.1304	0.1311	0.1478	0.1492	0.5067	0.5632	0.1893	0.1936	0.2564	0.2587
Fliers	80	87	90	107	209	241	142	139	200	221

A LQ means Lower Quartile. B UQ means Upper Quartile.

We evaluate the relative error of five items in the case of all features fed to the model. For every item, we consider two circumstances of applying log function and not applying log function. We calculate lower quartile, median, mean, upper quartile and number of fliers of each item under each circumstance (shown in Table 2).

The results indicate the applying of log function to the regression target can reach an overall reduction of relative error. The medians of all five items can reduce after we add the log function and the upper quartile can be restricted to a lower level as well. The relative error can reduce in 46 cases out of total 50 cases after we apply log function to the regression target. The results above reflect the accuracy promotion of applying log function.

Relative Error of Five Items

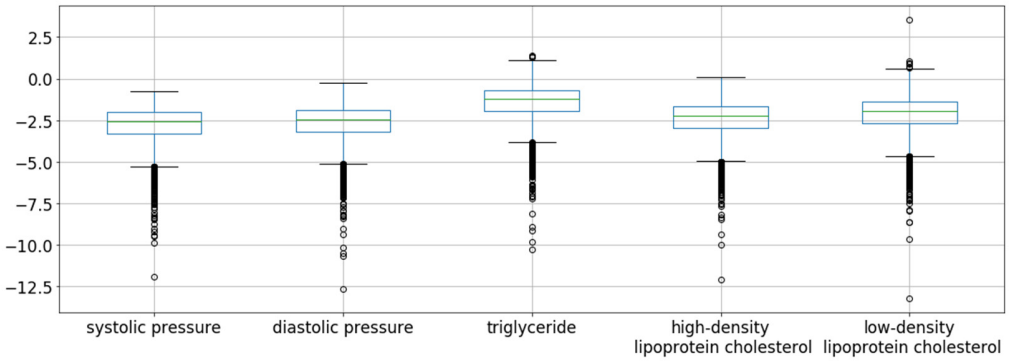

FIGURE 3. Boxplot of log relative error of five items

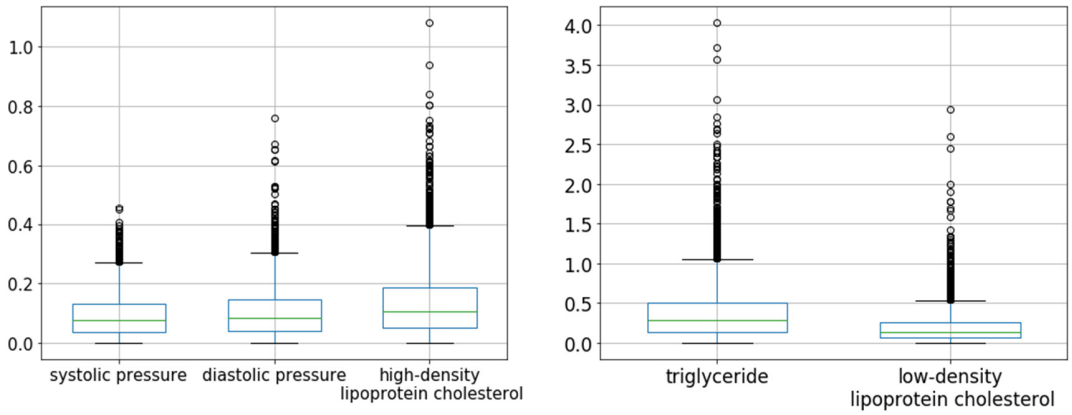

FIGURE 4. Boxplot of relative error of five items (after removing one flier up to 35.2 manually)

Figure 3 and 4 show the boxplots of relative error of five items. All five results have many outliers, and they are presented by hollow circles. To cover all the outliers in this picture 2, we have to apply log function to the relative error, because one test example's relative error of the LDL cholesterol prediction is 35.2 which is an obvious flier and much higher than others and it is presented as the only circle above 2.5 in figure 3. Triglyceride and LDL cholesterol's fliers range from 0.5 to 4.0, much wider than the other three items, which almost all distribute below 1.0. Thus, we have to plot them in two sub-figures.

The mean relative errors of systolic pressure, diastolic pressure, triglyceride, HDL cholesterol and LDL cholesterol are 0.0913, 0.1041, 0.3853, 0.1370 and 0.2040 respectively, while the medians are 0.0775, 0.0855, 0.2846, 0.1085 and 0.1413.

It means more than half patients can get a prediction of less than 20% relative error on five items except triglyceride and for all patients the relative error fades to almost 30%.

Among the five items, the results of systolic pressure, diastolic pressure and HDL cholesterol are better. Their outliers are much less than triglyceride and low-density lipoprotein cholesterol and they have lower relative errors. The upper quartile of four items except triglyceride is less than 0.3, while triglyceride's reaches 0.5067, which is not satisfactory. Nevertheless, for most examples, all five items are fitted well.

Comparison with Previous Studies

To evaluate cardiovascular risk in a more efficient and economical way, we propose a method based on the XGBoost algorithm. We can evaluate the patients' systolic pressure, diastolic pressure, triglyceride, HDL cholesterol and LDL cholesterol, which can be predictors as cardiovascular risk in subsequent work and we only use the medical examination records data. Compared with empirical method 1 2, our method is efficient and cheap, because we do not need experienced doctors or medical devices to directly detect patients' physical or chemical signals 3 4 5. What's more, we only use the easily accessible physical examination records to make predictions, so more personal information is not necessary 8. The experiment results of our method reflect a mean relative error range from 9.13% to 38.5% and the mean MSE of 0.0286.

CONCLUSION

In our research, we use a dataset of medical physical examination records, which are processed by NLP and transformed into digital form. We use five items systolic pressure, diastolic pressure, triglyceride, HDL cholesterol and LDL cholesterol as predictors to evaluate cardiovascular disease risk. The experiment results reflect a pretty high accuracy.

Our research only uses physical examination records data, which is easily accessible, and can reach a pretty high prediction accuracy. Our work can be helpful in conditions where patients' examination records have default items. Moreover, this method is cheap and efficient. So, it may be applied in hospital to evaluate the preliminary cardiovascular risk as an assisting for doctors to take subsequent examinant and prevention actions. The limitation is

that our result of triglyceride is not satisfactory, maybe raw data's information is not fully extracted. The future work includes extracting more information and improving the accuracy of results.

ACKNOWLEDGEMENTS

This work was supported by Shanghai Science and Technology Commission Project of China, No. 15DZ1202803 and No.17DZ1201605.

REFERENCES

1. Mistretta C A, Crummy A B. Diagnosis of Cardiovascular Disease by Digital Subtraction Angiography [J]. Science, 1981, 214(4522):761-765.
2. William J. Bommer, Larry Miller. Real-time two-dimensional color-flow Doppler: Enhanced Doppler flow imaging in the diagnosis of cardiovascular disease [J]. The American Journal of Cardiology, 1982, 49(4):944-944.
3. Matsushita K, Coresh J, Sang Y, et al. Estimated glomerular filtration rate and albuminuria for prediction of cardiovascular outcomes: a collaborative meta-analysis of individual participant data [J]. Lancet Diabetes & Endocrinology, 2015, 3(7):514-525.
4. Manninen V, Tenkanen L, Koskinen P, et al. Joint effects of serum triglyceride and LDL cholesterol and HDL cholesterol concentrations on coronary heart disease risk in the Helsinki Heart Study. Implications for treatment [J]. Circulation, 1992, 85(1):37-45.
5. Stamler J, Stamler R, Neaton J D. Blood pressure, systolic and diastolic, and cardiovascular risks: US population data [J]. Archives of internal medicine, 1993, 153(5): 598-615.
6. Staessen J A, Thijs L, Fagard R, et al. Predicting cardiovascular risk using conventional vs ambulatory blood pressure in older patients with systolic hypertension[J]. Jama, 1999, 282(6): 539-546.
7. Soni J, Ansari U, Sharma D, et al. Predictive data mining for medical diagnosis: An overview of heart disease prediction[J]. International Journal of Computer Applications, 2011, 17(8): 43-48
8. Wang C J, Li Y Q, Wang L, et al. Development and evaluation of a simple and effective prediction approach for identifying those at high risk of dyslipidemia in rural adult residents [J]. PLoS One, 2012, 7(8): e43834.
9. Chiu J P C, Nichols E. Named entity recognition with bidirectional LSTM-CNNs [J]. ArXiv: 1511.08308, 2015.
10. Alajali W, Zhou W, Wen S, et al. Intersection Traffic Prediction Using Decision Tree Models [J]. Symmetry, 2018, 10(9): 386.
11. Fan J, Yue W, Wu L, et al. Evaluation of SVM, ELM and four tree-based ensemble models for predicting daily reference evapotranspiration using limited meteorological data in different climates of China [J]. Agricultural and Forest Meteorology, 2018, 263: 225-241.
12. Szwabe A, Misiorek P. Decision Trees as Interpretable Bank Credit Scoring Models[C]//International Conference: Beyond Databases, Architectures and Structures. Springer, Cham, 2018: 207-219.
13. Friedman J H. Greedy Function Approximation: A Gradient Boosting Machine [J]. Annals of Statistics, 2001, 29(5):1189-1232.
14. Chen, Tianqi, and Carlos Guestrin. Xgboost: A scalable tree boosting system[C]. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016.
15. Chen T, Tong H, Benesty M, et al. xgboost: Extreme Gradient Boosting [J]. 2016.