Evaluation of the cardiovascular risk in middle-aged workers: an artificial neural networks-based approach

Alexander Sboev¹,²,³, Svetlana Gorokhova⁴, Viktor Pfaľ⁴, Ivan Moloshnikov¹, Dmitry Gudovskikh¹, Roman Rybka¹, Anton Selivanov¹ and Aleksey Serenko¹
¹ NRC «Kurchatov Institute», Moscow, Russia, sag111@mail.ru
² NRNU «MEPhI», Moscow, Russia
³ Plekhanov Russian University of Economics, Moscow, Russia
⁴ Research Clinical Center of Russian Railways, Moscow, Russia, cafedra2004@mail.ru

Abstract

A method of the evaluation of the risk of cardiovascular events in the group of middle-aged male workers was developed on the basis of artificial neural networks (ANN). The list of analyzed variables included parameters of allostatic load and signs of myocardial involvement. The results were compared with traditional scales and risk charts (SCORE, PROCAM, and Framingham). A better prognostic value of the proposed model was observed, which makes it reasonable to use both additional markers and ANN.

Keywords: data science, neural network, data analysis, cardiovascular risk

1 Introduction

The evaluation of cardiovascular risk is a difficult and important problem. The prediction of coronary heart disease (CHD) and major adverse cardiovascular events (MACE), e.g. myocardial infarction, sudden death etc. is especially significant in persons working with traffic safety. Risk charts, such as Systemic Coronary Risk Estimation (SCORE), PROCAM, Framingham risk score etc. are traditionally used in clinical practice to evaluate cardiovascular risk. The agreement between these risk scales is low (67% in average) (Allan, Nouri, Koronwyk, Kolber, Vandermeer, & McCormack, 2013), and their informative value significantly differs in various professional groups. These scales
also do not include markers of allostatic load as well as subclinical signs of the involvement of target organs. It is necessary to apply methods of mathematical data processing to extend the list of analyzed variables. Artificial neural networks (ANNs) can be used to solve the problem of the prediction of cardiovascular events (Atkov, et al., 2012) (Gorokhova, Sboev, Kukin, Rybka, Muraseeva, & Atkov, 2013). The aim of this study was to develop a prognostic model for CHD and MACE in the group of locomotive crew members (LCMs) on the basis of artificial neural networks. The paper presents an ANN-based model for the evaluation of cardiovascular risk in the group of middle-aged male workers.

2 Data Collection

The study included data on 106 LCMs (mean age 48.13, all men), who underwent a long-term medical observation. They were regularly subject to periodic and pre-trip ambulatory examinations, and in-depth hospital examination to diagnose CHD. The parameters of cardiovascular status were defined by a standard clinical examination, ECG, and echocardiography. Blood tests and stress tests (ECG and/or EchoCG) were performed in all the participants. The state of coronary arteries was assessed by coronary angiography and/or multi-slice spiral computed tomography (MSCT). CHD was diagnosed if documented signs of coronary myocardial ischemia, coronary atherosclerosis (according to coronary angiography and MSCT) or myocardial infarction were found. After that, patients were divided into two groups; CHD and No-CHD.

The initial set of variables in the database included over 60 signs:
— demographic data: age, length of service in the profession;
— medical history of hypertension, diabetes or other conditions;
— markers of allostatic load, such as body mass index (BMI), systolic and diastolic blood pressure (BP), hemoglobin, blood glucose, cholesterol, low density lipoprotein (LDL), triglycerides, creatinine etc.;
— signs of cardiovascular involvement, such as left ventricle mass index, end-diastolic diameter of left ventricle, arrhythmia, ECG QT-interval, sleep-disordered breathing, arterial stiffness, thickening of the carotid intima-media complex etc.;
— indexes of cardiovascular risk according to SCORE, PROCAM and Framingham.

3 Materials and Methods

The following steps were made to create the model:
1. pre-processing of input data with meaningful parameters extraction;
2. creation of the model of cardiovascular risk;
3. estimation of statistical accuracy of the newly developed and existing models on a dataset.

Tools used for problem solution were: Python 2.7.10 programming language; IDE: ipython notebook; software libraries: numpy 1.10.1, scipy 0.16.1, pandas 0.17.1, scikit-learn 0.17.0, matplotlib 1.5.1, seaborn 0.6.0.

3.1 Data Pre-processing

The first step of data pre-processing was data cleaning. The variables were excluded from the list if they:
1) lacked diagnostic value;
2) were obtained only in few workers (e.g. sleep apnea).
The next step was filling up the missing values of the remaining parameters with mean values. Then, data conversion was performed in order to improve the prediction accuracy of the classification model.

The pre-processing methods tested for the data conversion were input data normalization, principal component analysis (PCA), dimensionality reduction by choosing high-importance features (based on information entropy criterion), and PCA based on normalized data.

Normalization was performed with feature scaling algorithm, which brings all values of the parameters into the 0-1 range. PCA with 4 principal components was used. There were 6 high-importance parameters.

### 3.2 Classification Methods

All pre-processing methods were tested with all classification models. The following classification models were taken: support vector machine (SVM), decision tree classifier, extra-trees classifier, random forest classifier, adaBoost, bagging classifier, gradient tree boosting, probabilistic neural network (PNN).

Multiclassification algorithm for SVM in the current research was “one-vs-one”, i.e. classes were sequentially considered in pairs (Wu & Lin, 2004). SVM was based on a Gaussian radial basis function (RBF) kernel, which showed better results compared to linear or polynomial kernels. The testing of SVM with different class weights demonstrated that the optimal class weights were 1.0 for each class.

A decision tree was created based on the information entropy criterion (parameters with higher information entropy had higher importance in the classification problem). The size of the tree, the number of leaves and the number of objects in each node were not limited. The maximum entropy reduction was chosen as a splitting criterion.

Extra-trees classifier is a method based on building of a number random decision trees (a “forest”) (Breiman, 2001). Random forest classifier is based on a set of random decision trees for various subsamples of the input dataset (Geurts, Ernst, & Wehenkel, 2006). Bootstrap method was used for the division of samples. Parameters of decision trees in these methods were equal to the parameters of the decision tree classifier. There were 10 decision trees.

AdaBoost classification starts with fitting a classifier on the initial dataset and then fitting additional copies of the classifier into the same dataset, but with weights adjusted to incorrectly classified instances, in order to focus subsequent classifiers on difficult cases (Zhu, Zou, Rosset, & Hastie, 2009).

Gradient boosting classifier builds an additive model in a forward stage-wise fashion; it allows optimizing the arbitrary differentiable loss functions. In each stage, multiclass regression trees fit on the negative gradient of the binomial or multinomial deviance loss function (Friedman, 2001).

Scikit-learn software library implementation was used in most models; PNN was implemented in RBF kernel as a subclass of scikit-learn “base estimator” class, which was an abstract parent class for all classification models.

Optimal parameters of each classification model were fitted on the base of training and validation subsets of the original clinical data. After parameter fitting, each configuration of preprocessing method and classification model was tested.

The best classification results were obtained in extra-trees classifier in binary classification task, and in PNN in ternary classification task. Both models were based on normalized data pre-processed by PCA. The comparison and evaluation method of the models is described in the next section.

The estimation of the results of classification was performed with a stratified K-fold cross-validation of average results of 10 classification model executions. Configurations were estimated by ROC-AUC measure.
4 Results

The primary goal of binary classification in CHD prediction is to assess the risk of the disease; therefore, the main estimation parameter is ROC AUC of the first class (CHD group). Extra-trees classifier model based on normalized data pre-processed with PCA had the highest ROC AUC in the first class prediction (64%).

Ternary classification estimation with ROC AUC had two important parameters for CHD risk prediction: ROC AUC of the first class (CHD) and ROC AUC of the second class (MACE). The average value of these parameters was taken as the main estimation of configurations. The best model for the second-class prediction was the decision tree classifier based on information entropy criterion and normalized data pre-processed with PCA.

PNN model based on normalized data pre-processed with PCA had the highest average estimation (57%) and ROC AUC of the first-class prediction (63%). All top configurations for both binary and ternary classifications were based on normalized data pre-processed with PCA.

As a result, an optimal combination of methods to solve the problem of the prediction of heart diseases was determined. The best results were demonstrated in case of PNN with sigma = 0.1 and normalized data pre-processed with PCA with four principal components.

According to the ROC-analysis, the results of CHD prediction according to SCORE, PROCAM and Framingham scales were AUC = 0.72, 0.65, 0.69. However, AUC values for these risk calculators of myocardial infarction were significantly lower (0.34, 0.42, and 0.32, respectively).

5 Conclusion

SCORE, PROCAM, and Framingham scales of cardiovascular risk assessment are characterized with inconsistency of the results, and insufficient informative value to predict MACE (including myocardial infarction) in the group of middle-aged male railroad workers. This means that the use of these scales in this group of patients is limited. The best prognostic value of PNN model was achieved by the inclusion of the markers of allostatic load and signs of the left ventricular myocardial involvement (myocardial mass index). At the same time, the list of markers should not be excessive, and the variables should be thoroughly chosen. The results suggest that it is reasonable to use both additional markers and artificial neural networks.

References

Allan, G., Nouri, F., Koronwyk, C., Kolber, M., Vandermeer, B., & McCormack, J. (2013). Agreement Among Cardiovascular Disease Risk Calculators. Circulation, pp. 127:1948-1956.

Atkov, O., Gorokhova, S., Sboev, A., Generozov, E., Muraseeva, E., Moroshkina, S., et al. (2012). Coronary heart disease diagnosis by artificial neural networks including genetic polymorphisms and clinical parameters. J Cardiol., 190-194.

Breiman, L. (2001). Random Forests. Machine Learning, 5-32.

Friedman, J. (2001). Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics, 1189-1232.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. Machine Learning, 3-42.

Gorokhova, S., Sboev, A., Kukin, K., Rybka, R., Muraseeva, E., & Atkov, O. (2013). The Application of Artificial Neural Networks in the Diagnosis of Coronary Heart Disease. In: Network Topologies: Types, Performance Impact and Advantages/Disadvantages., 45-57.
Wu, T.-F., & Lin, C.-J. (2004). Probability Estimates for Multi-class Classification by Pairwise Coupling. *J Machine Learning Research.*, 975-1005.

Zhu, J., Zou, H., Rosset, S., & Hastie, T. (2009). Multi-class AdaBoost. *Stat Interface*, 349-360.