Features selection algorithm for an unbalanced sample of clinical data*

Karina Shakhgeldyan
Information Technology Department
Vladivostok State University of Economics
Vladivostok Russia
carinash@vvsu.ru

Boris Geltser
School of Biomedicine
Far Eastern Federal University
Vladivostok Russia
boris.geltser@vvsu.ru

Vladislav Rublev
School of Biomedicine
Far Eastern Federal University
Vladivostok Russia
dr.rublev.v@gmail.com

Basil Shirobokov
Information Technology Department
Vladivostok State University of Economics
Vladivostok Russia
wpn@inbox.ru

Dan Geltser
Imperial College London
London UK
dg1116@ic.ac.uk

Alexandra Kriger
School of Biomedicine
Far Eastern Federal University
Vladivostok Russia
kribalex@gmail.com

ABSTRACT

The aim of the study is to develop models of intrahospital mortality (IHM) prediction for coronary artery disease (CAD) patients after coronary artery bypass grafting (CABG) on an unbalanced sample.

Methods. Models for IHM prediction were built on the analysis of 866 electronic case histories of CAD patients, revascularized with the CABG operation. The patient cohort consisted of two groups. The first included 35 (4%) patients who died within the first 30 days after CABG, the second - 831 (96%) with a favorable operation outcome. We analyzed 99 factors, including the results of clinical, laboratory and instrumental studies obtained before CABG. For features compilation, classical filtering and model selection methods were used (wrapper method). The main problem with classical approach applying was an unbalanced sample, when one class contains only 4% of objects. In that case, it's not possible to apply the cross-validation procedure with three types of samples, standard quality metrics and multi-category factors.

Results. Features searching approach using the multi-stage selection procedure, which combined the validation of predefined predictors set, filtering methods and multifactor models development based on logistic regression, random forest (RF) and artificial neural networks (ANNs) was proposed. The models accuracy was evaluated by a combined quality metric. RF and ANNs based models allowed not only to build more accurate forecasting tools, but also to verify five additional IHM predictors.

CCS CONCEPTS

•Analysis methodology • Machine learning • Supervised learning by classification • Feature selection

KEYWORDS

Wrapper features selection method, filter features selection method, evaluation metrics, unbalanced sampling

1 Introduction

Machine learning (ML) methods have been used increasingly among clinical trials over recent years. Their objective is the models development of output predicted variables calculation, based on several input factors, characterizing the clinical and functional status of patients with various diseases, pharmacotherapy or surgical intervention options, and other characteristics [1,2]. Clinical medicine refers to the specific field of knowledge and practice where special significance is attached to the evidences of predictive potential of the prognostic models factors.

There are three main groups of methods for feature selection: filter methods, wrapper methods and embedded methods [3-5]. Filtration methods include statistical analysis of intergroup

*The reported study was founded by RFBR according to the research projects 18-29-03131 and 19-29-01077

†Author Footnote to be captured as Author Note

WOODSTOCK'18, June, 2018, El Paso, Texas USA

differences, correlation assessment, variance analysis, which all allow us to verify the linear relationships intensity and individual factors and predicted variables associations. At the same time, the models development makes it possible to take into account inter-factor relationships, non-linear associations between predictors and the predicted variable.

In clinical medicine, predictors should be identified, validated and justified. Nevertheless, in many cases, the data sets on which the study is performed are unbalanced and one class contains less than 10% of the analyzed objects. An example of such tasks is intrahospital mortality after heart surgery.

Coronary artery disease (CAD) is one of the main mortality causes worldwide. Among cardiovascular diseases, it makes up more than 50%. Coronary artery bypass surgery (CABG) is one of the most common coronary blood flow repair procedures among CAD patients. In this regard, interest in studying the role of factors affecting the risk of adverse CABG outcomes is constantly growing [6]. The level of which among patients under 70 years old is around 1-3%, and among people over 70 years old up to 6% [7]. Thus, even with sufficiently large sample of data, two thousand patients for example, the group of patients with an unfavorable outcome would be only 20-60 people. In that case, three subsamples: training, test and validating, contains small data in the unfavorable outcome group, which leads to a decrease in the reliability of the assessment results. The main metrics that are traditionally used to assess the quality of models: accuracy, precision, area under the ROC curve (AUC) in the case of unbalanced data can be biased. Categorical variables with more than two values could be incorrectly estimated due to the insufficient number of objects in one of the groups. These problems of unbalanced samples require procedures that are more meticulous for the found predictors proof.

2 Materials and methods

We conducted a retrospective analysis of 866 electronic case histories (ECH) of CAD patients, operated in 2008-2018 by isolated CABG in “Primorsky Regional Clinical Hospital №1” Vladivostok city. ECH were converted to a dataset by automatically extracting data from the medical information system. The missing individual indicators values were supplemented by information obtained from the archive of medical records on paper. Missing data that could not be filled out were excluded from the analysis. Thus, the size of the dataset for individual parameters ranged from 687 to 866 ECH.

The examined cohort of patients consisted of two groups. The first included 35 (4%) patients who died within the first 30 days after CABG, the second - 831 (96%) with a favorable operation outcome, indicating an unbalanced sample. The features search was carried out among 99 factors characterizing the clinical, medical history and functional status of each patient before CABG. These factors included the results of clinical, laboratory and instrumental studies.

Searching for features from a large number of factors can be started with some basic version of predictors, if they were highlighted in previous studies. Over recent years, unified tools for adverse events prediction, based on the results of large prospective studies have been increasingly used in clinical cardiology. So, in cardiology, well-known scales are SCORE, ASSIGN SCORE, predicting the risk of death from cardiovascular disease in the next 10 years, EuroSCORE and EuroSCORE II, estimating probability IHM within 30 days after CABG [8]. Scales EuroSCORE and EuroSCORE II are logistic regression (LR) models and include 16-18 features, characterizing clinical and functional status of patients before surgery, types and urgency of cardiac surgery [9]. These predictors can be considered as a basic set that must be verified on the analyzed cohort, and then expanded because of the study with new factors.

3 Results

A feature of clinical research is the same datasets using for solving various problems, for each of which a certain set of ECH could be selected. In this case, part of the factors may be a constant. For example, from data describing patients who underwent various heart surgeries, the sample of those who underwent only CABG contains a constant factor characterizing the type of surgery. Some factors may also have a low incidence, below 1%. If in an unbalanced dataset all the few cases organized a group containing the majority of patients, then such variable will not have an informative potential. The analysis of works on existing scales showed that they often contain factors, not all of which are of statistical significance (p-value > 0,05), which decreasing their prognostic value. In addition, sometimes factors that correlate with each other are included in the scales, which create the problem of multicollinearity [10]. Since most of the existing scales in clinical medicine are LR, nonlinear relationships remain unaccounted. Thus, our approach should include procedures for filtering factors, accounting for nonlinear associations, and working on an unbalanced sample with a small number of objects in one of them.

3.1 Features search algorithm for unbalanced clinical data

Search approach for predictors among large number of analyzed factors for prognostic modeling of after surgery survival contains the following steps.

1. Choosing a common set of factors $P=\{p_i,\ i=1,..,N\}$, describing the patients who underwent surgical intervention, we define or calculate the dependent variable as a dichotomous category. All variables in continuous form are left without conversion to categorical.
2. We calculate various indices characterizing the clinical condition of patients, which a priori may have predictive potential.
3. Using the filtering method of searching for factors with constant values, we remove factors that are constants from the P set on the analyzed dataset. $P^{f1}=P\oplus P^k$, where P^k – indicators variety representing a constant.

4. Using filtering method for the search of low-variant variables, delete factors from P^{f1}, that have a low frequency of occurrence (<1%) P^d, while falling into a group with a large number ECH: $P^{f2}=P^{f1}\oplus P^d=P\oplus P^k \oplus P^d$.
5. We select previously developed scales for assessing the probability of the analyzed event occurrence. In clinical medicine, such scales are LR models.
6. Complement the P^{f2} set with categorical factors analogous to some continuous factors: $P^{full}= P^{f2}\cup P^{c}$. The criteria for dividing a continuous variable into categorical values are selected from known scales (see above) and ratios used in clinical medicine. Further univariate analysis involves the study of both continuous and their corresponding categorical variables.
7. Analyzing P^{full} factors using filtering methods (checking intergroup differences – Student, Mann-Whitney, Fisher tests и χ^2) and selecting those that showed statistically significant differences (p-value < 0,05) $P' \subseteq P^{full}$.
8. Using single-factor LR models, binding P^{full} factors and the dependent variable, we form a factors subset $P^o \subseteq P^{full}$, which are statistically significant in univariate LR-models.
9. We form a set $P^{filter} = P' \cup P^o$, which combines factors that have linear and fairly explicit relationships with the dependent variable.
10. Estimating the probability of an event occurrence in the study cohort according to the known scales. Assessment is performed using three metrics: sensitivity, specificity and AUC. All subsequent models are compared with this rating and should only be considered if they improve it.
11. From the P^{filter} set we select predictors P^s, which are used in previously developed scales, and evaluate them in terms of multicollenarity problem. If two factors are linearly dependent, one of them is either not considered or replaced by a similar factor that does not have a linear relationship with significant scale predictors. P^s includes, along with continuous, their categorical analogues.
12. Wrapper methods based procedure starts begin with LR. Based on the P^s predictors, the F^s model is built on the entire data sample. For further consideration in multifactor LR models, only those predictors are selected (P^{s1}), that statistically significant throughout analyzed cohort. From the continuous and categorical variants of one factor, we select those with greater statistical significance. In unbalanced samples, this probability is higher for continuous factors.
13. Cross validation procedure of LR model with P^{s1} predictors we perform on the training and evaluate on a test sample by three quality metrics (sensitivity, specificity and AUC). If quality metrics improved over existing scales, then the model $F^s(P^{s1})$ considered basic and further comparisons are made with it. The F^s model contains only those predictors that have linear associations with the predicted event.
14. For further analysis, we leave in the Pf^{ilter} set predictors those variants of continuous indicators that are included in P^{s1}.
15. To extract factors that nonlinearly affect the predicted event, we use machine learning methods, for example RF and ANNs. The best model parameters finding cycle begins with a basic set of predictors P^{s1}. After fixated P^{s1}, we look for such model parameters that will ensure accuracy higher than LR on test samples when performing cross-validation. Due to sample imbalance, generalized metric appliance, for example AUC, does not provide a correct model quality assessment, therefore, we proposed to use unbiased metrics - Matthews correlation coefficient or the average between sensitivity and specificity for example.
16. Due to small number of objects in one of the classificatied groups among many problems of clinical medicine, it is not possible to use the third validating sample to select the best model after the cross-validation procedure. It is proposed to choose the best among models within cross-validation by averaging two quality metrics (sensitivity and specificity) and quality checking on test and on a full data sample. We consider the worst of these two results to be the quality assessment of the best model. We add one by one factors from P^{full} to P^{s1}, at the same time we start with the factors P^{filter} included, and performing the selection procedure for model parameters and cross-validation to a new set of predictors. If factor does not improve the quality of the model, then we move it into the set P^R, for further analysis. For ANNs models there are infinitely many different options for parameters, since the model quality with an unbalanced sample among a small number of objects in one of the groups is affected not only by the network architecture, but also by a random number. That is why we propose to focus not only on averaged metrics during cross-validation, but also on the best models selected inside this procedure.
17. At the previous step, all factors independently affecting the predicted variable were identified. We repeat the previous step, while adding from the set P^R 2 or more factors at the same time, combining factors from one semantic group into a stack. For example, factors responsible for lesions of target organs or indicators of blood coagulation.
18. The predictors, which are included in to majority of the best models developed by various methods, were considered the most influential in the probability of an unfavorable outcome.

3.2 Searching for features predicting IHM after CABG

For the prediction of IHM after CABG via filtration methods, we removed 14 out of 113 considered indicators, which were either constant in our dataset of patients with CAD who underwent CABG surgery, or had a very small variance. The results of the static analysis showed that only certain factors had significant intergroup differences. Thus, from 18 EuroSCORE II scale indicators, statistically significant intergroup differences were recorded only in 7 parameters, which included age, left ventricle (LV) ejection fraction (EF) less than 30%, LVEF from 30% to 50%, previously transferred myocardial infarction (MI), atherosclerotic lesion of the peripheral arterial pools, urgency of CABG procedure, congestive heart failure (CHF) III-IV functional classes (FC). Such factors of EuroSCORE II as gender, creatinine clearance, mean pulmonary artery pressure (mPAP), any form of angina, presence of chronic obstructive pulmonary disease (COPD), type 2 diabetes (T2D) did not have significant intergroup differences. Additional factors with intergroup differences were identified: systolic blood pressure (SBP), signs of hypertension (HTN) and aortic stenosis.

To assess the individual risk factors influence on IHM, we constructed one-factor LR models with the weight coefficients calculation, characterizing the predictive value of the analyzed indicators (Table 2). We considered continuous factors, previously normalized, along with their categorical counterparts, since in our study continuous factors had greater statistical significance. For example, LVEF in the continuous case had p-value=0,000026, while LVEF <30% was significant at p-value=0,0057.

As an analysis result, it was found that LVEF has the maximum effect on IHM in continuous form (-4,52). Comparable variables are: SBP (-4,11) and age in continuous form (3,6). A slightly lesser effect on IHM has HR (2,8), LVEF < 30% and serum creatinine level (2,4). The least impacts that have statistical significance are exerted by the following indicators: age in continuous form with a value of more than 60 years (1,8), recently transferred MI (1,7), operation urgency (1,67), CHF III-IV FC (1,6), aortic stenosis (1,5), HTN (-1,2), LVEF from 30 to 50% (1,1), as well as atherosclerotic lesions in the peripheral arteries (0,93).

Two variants of multifactor LR models were developed (Table 1). In the first one 7 statistically significant EuroSCORE II scale predictors were used, and in the second - they were supplemented with 3 new factors (heart rate, signs of hypertension and aortic stenosis). It was found that predictors in the form of continuous variables (age, serum creatinine and LVEF) in relation to their categorical counterparts, which are used in the scale EuroSCORE II, possessed the best prognostic properties, as evidenced by the weight coefficients of the one-factor model and the quality metrics of multifactor models. At the same time, such EuroSCORE II predictors, as: gender, CHF II FC, stable angina IV FC, COPD, T2D, mPAP in the author multivariate models were also statistically insignificant, as they were in the univariate ones. On the contrary, recent MI factor, whose predictive value in EuroSCORE II scale was insufficient (p>0,05), in the author's multifactor model has acquired significant prognostic value (p<0,0001). It should also be noted that EuroSCORE II scale used a combination of linearly dependent indicators of age and creatinine clearance, which created a problem of multicollinearity, limiting the prognostic effectiveness of the scale. Creatinine clearance is calculated using creatinine, age, and weight, and naturally correlates with age. To solve this problem, the author's LR model used linearly independent predictors (age and serum creatinine concentration), which made it possible to increase the accuracy of the forecast.

Table 1.

Features weights for multivariate LR models of intrahospital mortality prediction after CABG

Predictors	Authors LR model with EuroSCORE II predictors		Authors LR model with EuroSCORE II predictors and add-ons	
	Coefficient	p-value	Coefficient	p-value
Age	2,58	0,04	3,28	0,012
Serum creatinine	2,64	0,06	2,85	0,058
Peripheral artery disease	0,89	0,023	1	0,014
Recently transferred MI	1,77	<0,0001	2,1	<0,0001
LVEF	-3,02	0,016	-2,78	0,032
Urgency operation	1,64	0,03	1,48	0,034
CHF III-IV FC	1,69	<0,0001	1,72	0,000085
HR*	-	-	3,69	0,0089
Aortic stenosis *	-	-	1,66	0,01
HTN*	-	-	-1,62	0,0048
Constant	-5,66	<0,0001	-6,25	0,000002

To assess the accuracy of the models, we used three quality metrics: AUC, sensitivity and specificity. The criterion threshold for assessing the probability of IHM, according to the EuroSCORE II scale is 5%. This means that with a predicted EuroSCORE II probability above 5%, the patient is at a very high risk of IHM after CABG. For further analysis, the basic IHM probabilities were calculated on patients cohort using the EuroSCORE II scale and predictors. It was found that the forecast quality for AUC metrics, sensitivity and specificity for the "classic" EuroSCORE II scale with 18 predictors was 0,73, 0,25 and 0,92 respectively (Table 2). This indicates insufficient accuracy during testing on the analyzed cohort and further improvement needed.

At the next analysis step IHM models were developed based on the LR, RF, and ANNs methods using the cross-validation procedure, which was provided by multiple (at least 500 times) random division of the examined cohort into training and test samples in the ratio of 75% and 25%, respectively. In the training samples the models parameters were selected so that the averaged values of the three quality metrics in the corresponding test samples reached maximum values. For LR model parameters were selected predictors that had statistical significance in a multivariate model.

The results of analysis presented by two groups of quality metrics. The first of them is averaged values of two metrics on test samples randomly generated from the examined cohort. Selected predictors are common for the constructed models, some of which were not previously included in any of the IHM rating systems after CABG.

Table 2

Accuracy evaluation of the developed prediction models for hospital mortality after CABG

	Metrics under cross-validation			Metrics of the best model		
	AUC	Sensiti	Specif	AU	Sens	Specific

		vity	icity	C	itivity	ity
LR-0	0,75	0,25	0,92	0,75	0,25	0,92
LR-I	0,83	0,74	0,78	0,86	0,83	0,76
LR-II	0,85	0,7	0,8	0,89	0,83	0,78
RF-I	0,71	0,69	0,71	0,89	1	0,79
RF-II	0,78	0,82	0,74	0,9	1	0,81
ANNs-I	0,85	0,8	0,9	0,96	0,93	0,91
ANNs-II	0,9	0,86	0,94	0,999	1	0,998
ANNs-III	0,95	0,92	0,98	1	1	1

4. Discussion

A comparative analysis showed significant differences between the EuroSCORE II models (LR EuroSCORE II on 18 predictors) and the author's multivariate LR models. The author's LR-I model, built on 7 statistically significant predictors of EuroSCORE II, provided an sensitivity (0.83) and AUC (0.86) indices increasing on the test sample against the background of lower specificity (0.76). The expanded LR-II model with additional parameters of heart rate, hypertension and aortic stenosis sign allowed increasing of specificity indices up to 0.78 and AUC up to 0.89. The best RF-I model with EuroSCORE II predictors showed 100% sensitivity (versus 0.83 for LR-I) with a comparable level of specificity (0.76 vs 0.79) and AUC (0.86 vs 0.89). The RF-II model with three additional predictors had better quality, on average and as the best model choosing, against RF-I and all LR models. Three ANNs showed better results than the LR and RF models, both on average and as the best models. The addition of 3 predictors (heart rate, SBP and aortic stenosis sign) significantly increased the quality of the ANNs-II model compared to ANNs-I (7 main predictors). Two additional factors characterizing myocardial hypertrophy (RWT and LVMI) allowed to obtain the best ANNs-III model with 100% accuracy.

ANNs models architecture during the study varied from 1 to 200 layers and from 3 to 100 neurons in each layer. The final version included 3 hidden layers with 7 or 10 neurons in each. The sigmoid suited the best as an activation function on all three layers. Optimization was introduced by the Adam algorithm with an accelerated Nesterov gradient. Main parameters of the developed RF models were 1000 trees for "voting" and 6 to 8 splitting signs.

4. Conclusion

The study allowed us to develop a feature selection algorithm for IHM prediction after CABG and build models with high predictive accuracy. Models based on ML (RF and ANNs) methods showed high accuracy and allowed us to identify several additional IHM predictors after CABG.

ACKNOWLEDGMENTS

The reported study was founded by RFBR according to the research projects 18-29-03131 and 19-29-01077.

REFERENCES

[1] MN Karim, CM Reid, M Huq, et al. (2018). Predicting long-term survival after coronary artery bypass graft surgery. Interactive Cardiovascular Thoracic Surgery, 26(2).

[2] I Kurt, M Ture, AT Kurum (2008). Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. Expert Systems with Applications, 34(1), 366-374.

[3] I Guyon, A Elisseeff (2003). An Introduction to Variable and Feature Selection. Journal of Machine Learning Research, 3, 1157-1182.

[4] Y Mao, Y Yang (2019). A wrapper feature subset selection method based on randomized search and mulripayer structure. BoiMed Research International, 2019, 9. Doi: 10.1155/2019/9864213

[5] N Sranchez-Maro˜no, A Alonso-Betanzos, RM Calvo-Estrevez (2009) A Wrapper Method for Feature Selection in Multiple Classes Datasets IWANN, Part I, LNCS 5517, pp. 456–463, 2009.

[6] H Yamaoka, K Kuwaki, H Inaba, T Yamamoto, TS Kato, S Dohi, et al. (2015) Comparison of modern risk scores in predicting operative mortality for patients undergoing aortic valve replacement for aortic stenosis. J. Cardiol, PII: S0914-5087(15) 00282-8.

[7] MP Rao, SM Al-Khatib, SD Pokorney, et al. (2017). Sudden Cardiac Death in Patients With Ischemic Heart Failure Undergoing Coronary Artery Bypass Grafting: Results From the STICH Randomized Clinical Trial (Surgical Treatment for Ischemic Heart Failure). Circulation, 135 (12).

[8] A Sedaghat, JM Sinning, M Vasa-Nicotera, A Ghanem, C Hammerstingl, E Grube et al. (2013) The revised EuroSCORE II for the prediction of mortality in patients undergoing transcatheter aortic valve implantation. Clin. Res. Cardiol, 102 (11), 821–9.

[9] SAM Nashef, F Roques, LD Sharples, J Nilsson, C Smith, AR. Goldstone, U Lockowandt (2012). EuroSCORE II. European Journal of Cardio-Thoracic Surgery, 2012, 41 (4).

[10] J Faraway. 2016. *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models* (2nd ed.). Chapman & Hall/CRC Texts in Statistical Science Book, Kindle Edition.