Linear Bayes Classification for Mortality Prediction

Martin Macaš, Jakub Kuzilek, Tadeáš Odstrčilík, Michal Huptych

Czech Technical University, Prague, Czech Republic

Abstract

The paper describes our solution for PhysioNet challenge 2012, which finally achieved 4th place in Event 1 and 3rd place in Event 2. To predict mortality of ICU patients, we used simple linear Bayes classifier, for which we selected features using Social Impact Theory based Optimizer.

1. Introduction

This paper describes our solution for PhysioNet challenge 2012 [1]. Its main aim is to predict mortality of ICU patients and obtain high S_1 and small S_2. S_1 is defined as the smaller one of sensitivity and positive predictivity and S_2 is the Hosmer-Lemeshow H statistic [1]. Each record can be understood as consisting of 37 time series of different lengths, each corresponding to one variable measured during the patient's stay at ICU. For each person, we extracted high number of features. The original feature set was further reduced using some preprocessing and selection and used to train linear Bayes classifier.

2. Methods

2.1. Feature extraction and preprocessing

Two types of features were used. In the left column of upper part of Table 1, one can see features that are general features related directly to the patient. Most of these features are outputs from the different standard scoring systems [1], the others are Age, Gender, Height and ICU type. The second type of features are related to the different variables measured during patients stay at ICU. For each such variable, these features are extracted from the corresponding time series. These features are summarized in the left column of lower part of Table 1. If the feature value is incalculable (for example, if the variable was not measured), it is replaced by Not-a-Number (NaN) value.

There was no feature scaling needed, because the classifier is invariant to a linear feature scaling. The scaling was used only for some additional experiments with neural networks and other classifiers that are not described here. For some settings and experiments we also examined potential utility of outlier filtering, however a removal of outliers did not lead to some improvements. Some preliminary experiments with Principal Component Analysis also did not bring any significant improvements for our solution. Therefore, the only preprocessing that was performed for all experiments was the elimination of features with many NaN values – features with more than 200 NaN values were removed. This strong elimination of features reduced the dimensionality from 935 to 352 features. The remaining NaN entries were replaced by mean values of their features. In some cases, mostly in latter experiments, we also performed correlation analysis (Correlation in Table 2). First, correlation coefficients were computed for each pair of features and from each pair with the absolute value of the coefficient grater than 0.6, one feature was eliminated. This type of preprocessing, if used, lead to a further dimensionality reduction to 95 features.

2.2. Classification and prediction

We employ a common feature-based approach. After the extraction, preprocessing and selection, the features are used for building the classifier. The underlying parametric classification model is a Bayes classifier [2]. It uses Bayes theorem to compute posterior probability of each class with label l from likelihood $p(\mathbf{x}|l)$ and class prior probability $p(l)$:

$$p(l|\mathbf{x}) = \frac{p(\mathbf{x}|l)p(l)}{p(\mathbf{x})}. \quad (1)$$

The feature vector $\mathbf{x}$ is further assigned into the class that maximizes posterior probability:

$$CLASSIFICATION(\mathbf{x}) = \arg\max_{l\in\{0,1\}} p(l|\mathbf{x}). \quad (2)$$

A particular case of Bayes classifiers is the linear Bayes classifier, which assumes Gaussian class–conditioned distributions with the same covariance matrix for both classes which leads to a linear decision boundary. The expectation vector is estimated from training data using the sample mean. The covariance is estimated using the sample

Table 1. The list of all features. The upper part summarizes 10 general features related to the patient. For each of 37 variables measured during patient's hospitalization at ICU, each of 25 features listed in the lower part of the table were computed. Totally, we extracted 935 features for each record. Right column lists variables for which the feature was selected in Entry 8.

Feature description	Selection in Entry 8
Age	
Gender	
Height	
ICU type	✓
SOFA score	
SAPS I score	
SAPS II score	✓
Apache I score	✓
Apache II score	
Apache III score	✓
Apache IV score	
1 if all derivatives of the feature are non-zero	HCO3,HR
difference between first and final value	HCO3,HR,Temp,WBC
first value	BUN,GCS,HCO3,MG,Urine
kurtosis	Platelets,WBC
maximum derivative	BUN, GCS, HCO3
difference between maximum and minimum derivative	HR,Temp,Urine
maximum value	HR,Temp,Weight
mean derivative	BUN,GCS,HCO3
mean value	GCS,Glucose,Na,Weight
absolute difference between median and mean value	GCS,HCO3,Mg,Na,Platelets
median of the derivative	BUN,Platelets
median value	BUN, Creatinine, GCS, K
minimum value	GCS,HCT,Mg,Platelets,Weight
mode, or most frequent value	HCT,HR,Mg,Temp
number values measured	ALT, AST, BUN, Bilirubin, Cholesterol, Creatinine, Glucose, HR, K, MechVent, Mg, NIDiasABP, Platelets, Urine, WBC, Weight
lower quartile	Creatinine,HCO3,HCT,HR,Temp,Urine,Weight
upper quartile	BUN, GCS, Glucose, Mg,Temp,
difference between maximum and minimum value	Creatinine,K,Na,WBC
signum of the mean derivetive	Urine
standard deviation of the derivative	BUN,Creatinine,HCT
standard deviation	Glucose,K,Mg,Temp,Urine
sum of values	BUN,Na,Platelets,Weight
trend (slope of a line fitted to values)	HR,Na,Platelets,Urine
variance	BUN,GCS,HR,Mg,WBC
variance of derivative	Creatinine,Temp,WBC

covariance matrix. Finally, linear Bayes classifier is used for computation of class posteriors. We also experimented with other classification models, however the linear Bayes clearly outperformed all the others in some preliminary experiments.

For our classification problem, we do not have any exact knowledge about the prior probabilities $p(l)$, but we can assume that prior probability for positive class (in-hospital death) will be lower than 0.5. Therefore, we used the prior $p(1)$ as a tuning parameter.

The in-hospital mortality risk was predicted by the corresponding posterior for $l = 1$ multiplied by a pre-tuned constant:

$$RISK(\mathbf{x}) = \alpha p(1|\mathbf{x}), \tag{3}$$

where α is the second tunable parameter of our approach.

2.3. Feature subset selection

A feature selection process usually consists of two main components - a search criterion, which evaluates potential feature subsets, and a search method, which seeks for a minimum of the criterion. Here, we use the wrapper approach to feature selection and our criterion value depends on the Bayes classifier. In Phase 2, the criterion was minimized using the SSITO method. In all experiments of Phase 2, our particular feature selection criterion was a linear combination of estimated score for Event 1 (S_1) and score for Event 2 (S_2):

$$f = -\omega_1 S_1 + \omega_2 S_2, \tag{4}$$

where ω_1 and ω_2 are weights for particular events (see

Table 2. Summary of particular entries. CV means 10-fold crossvalidation estimate of scores on set A dataset.

#	Classifier	Pre-selection	Search	Estimate	Criterion	Dim	CV 1	CV 2	S_1 on set B	S_2 on set B
1	Linear	None	RANK	RES	Error	10			0.40	35.8
2	Linear	None	RANK	RES	Error	10			0.40	30.0
3	Quadratic	None	RANK	RES	Error	10			0.30	67.4
4	Quadratic	None	RANK	RES	Error	10			0.34	61.0
6	Quadratic	None	SSITO	RES	$-S_1 + 0.0005S_2$	162			0.25	36892
7	Linear	Correlation	SSITO	10CV	$-S_1 + 0.003S_2$	41	0.44	19.2	0.45	NaN
8	Linear	None	SSITO	2CV	$-S_1$	110	0.47	24.1	0.47	12.8
9	Linear	Correlation	SSITO	10CV	$-S_1$	56	0.45	17.1	0.47	22.9
10	Linear	Rank/in-in distances	SSITO	10CV	$-S1$	9			0.44	36.1

6th column of Table 2). It was observed that a maximization of S_1 leads to small S_2, although minimization of S_2 does not lead to high values of S_1. Therefore we mostly preferred S_1 in criterion. It is crucial for our solution, that S_1 and S_2 values should be estimated using a stratified cross-validation technique. For example, Entry 6 was optimized using re-substitution method (training and testing on the same whole set A). Especially for the quadratic classifier, the feature selection leads to a strong overfitting. Although extremely promising criterion value was reached for set A ($S_1 = 0.854$ and $S_2 = 3.651$), set B scores for Entry 6 were relatively bad. This phenomenon was also recognized for linear classifier (although more weakly) during some preliminary experiments. Thus, we decided to use the 2-fold or 10-fold cross-validation (2CV and 10CV in Table 2) estimates and computed the S_1 and S_2 values by averaging over multiple cross-validation splits.

The optimization approach used for feature selection was Simplified Social Impact Theory based Optimizer (SSITO) [3, 4]. It tries to take a model from social psychology, adapt it, and use it in the area of parameter optimization. It is an attempt to use simulated people to make a decisions about solutions of an optimization problem. The simulation is based on simple opinion formation models widely used in computational psychology. It is a novel population-based optimization methods, in which the candidate solutions influence each other and try to converge into a "good" consensus.

3. Results

We performed a huge number of experiments to be able to find a good classification system. The main purpose of the experiments was to answer many questions related to particular components of our system.

3.1. Phase 1 solutions

There was a difference between Phase 1 and Phase 2 of the challenge. A schematic diagrams are depicted in Figure 1. During the Phase 1 (upper part of Figure 1), we did not use any special dimensionality reduction except the filtering of features with too many NaNs. Further we evaluated the 352 features one-by-one individually using resubstitution error computed for 1-dimensional Bayes classifier trained and tested by the examined feature on the same data. Further we selected only 10 features with best evaluation and used them directly to create final classifier. Uniform prior probabilities were used without any tuning. Parameter α influences only the value of Score 2 and there was only one minimum of Score 2 for $\alpha = 0.45$, which was used for risk prediction. This very simple system was used with linear classifier for the first two entries (see second column of Table 2). Further, we observed that the use of quadratic classifier leads to much worse results on the set B. This phenomenon was observed also in latter experiments.

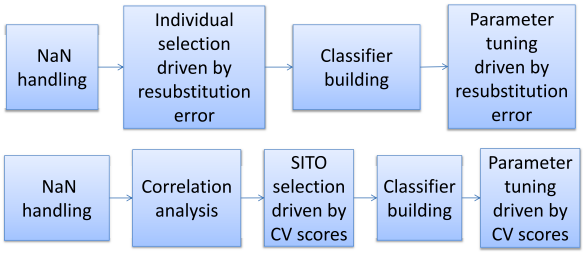

Figure 1. The two approaches to the solution. The first approach, depicted in the upper part, was used for all Phase 1 experiments and challenge entries. The second approach, depicted in the lower part, was used for most experiments of the Phase 2. An exception is the Entry number 6, where resubstitution estimate of scores was used.

3.2. Phase 2 solutions

In the Phase 2, we focused on the feature selection based on the SSITO method. For Entry 6, it was found that SSITO method is able to find extremely good values of the cost function ($S_1 = 0.854$ and $S_2 = 3.651$) if the quadratic classifier was used. However, the poor result on set B (see Table 2) indicated that there is probably a strong optimistic bias in the resubstitution estimate. This also corresponds

to Figure 2, where a significant optimistic bias of quadratic classifier with resubstitution is also evident. Therefore, we used the stratified crossvalidation in all remaining entries of Phase 2 to estimate the true score values.

To facilitate the optimization process, we also tried to apply pre-selection, which preceded the main SSITO optimization process. The first type of pre-selection is a filtration of correlated features. Entries 7 and 8 used the correlation analysis. In most experiments, it lead to much faster search in 95-dimensional space, however it did not help the SSITO optimizer to find a better result. The second type of pre-selection similar to that described in previous section for Phase 1 solutions was the individual evaluation of features and using only the best 20 features. It was used only in Entry 10 and did not lead to any special improvement.

Further, it is important to mention the number of folds in cross-validation. The best result was reached when the two-fold cross validation was used for the computation of the selection criterion. This corresponds to the results described in [5], where the repeated 2CV was reported to lead to the best stability and performance of wrapper methods. In [4], we also observed the superiority of the two-fold setting. This result is probably due to the larger testing dataset in the two-fold setting, which leads to smaller variance of the estimate.

Finally, we can summarize our solution. Entry 8 performed best on set-B and thus it was selected for the final evaluation on set-C. It simply used linear Bayes classifier and SSITO method guided by two-fold cross validation estimate of true score for Event 1. It is interesting that it simultaneously lead to a very good value of score for Event 2. The list of variables for which each feature was selected by SSITO method for Entry 8 is depicted in the right column of Table 1.

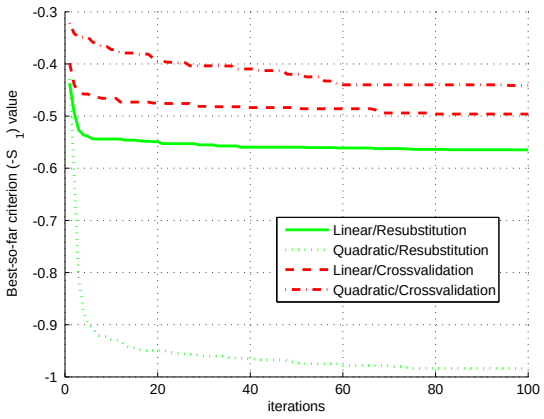

Figure 2. An example of optimization for different classifiers and feature selection criteria

4. Conclusions

In the preliminary evaluation on set B, our Entry 8 achieved $S_1 = 0.475$, which is the 8th best set B solution, and $S_2 = 12.820$, which is the 1st best set B solution. However, in the final evaluation of the challenge entries on set C, out Entry 8 achieved score for Event 1 0.4928 which is the 4th best solution and score for Event 2 0.247, which is the 3rd best solution. Many interesting conclusions can be drawn from many experiments (some of them were not described in the paper due to the space limitation). Linear classifier mostly outperformed the linear classifier regarding both the set B testing and the crossvalidation. SSITO method was used, because it significantly outperformed some other algorithms - sequential forward search, sequential forward floating selection, or Particle Swarm Optimization in some preliminary experiments. SSITO method was able to significantly improve the results in Phase 2.

Acknowledgements

The research was supported by ČVUT Grant no. SGS10/279/OHK3/3T/13.

References

[1] Silva I, Moody GB, Scott DJ, Celi LA, Mark RG. Predicting mortality of patients in intensive care: The physionet/computing in cardiology challenge 2012. In CINC 2012, volume 39. 2012; .

[2] Duda RO, Hart PE, Stork DG. Pattern Classification. Wiley, 2001.

[3] Macaš M, Lhotská L. Optimizers derived from human opinion formation. In Proceedings of Third World Congress on Nature and Biologically Inspired Computing, NABIC2011. 2011; .

[4] Macaš M. Opinion formation inspired search strategies for feature selection. Ph.D. thesis, Czech Technical University in Prague, 2012.

[5] Křížek P. Feature selection: stability, algorithms, and evaluation. Ph.D. thesis, Czech Technical University in Prague, 2008.

Address for correspondence:

Martin Macaš , Dpt. of Cybernetics,CTU-FEE,
Karlovo namesti 13, 121 35 Prague 2, Czech Republic
macas.martin@fel.cvut.cz