LASSObased falsepositive selection for classimbalanced data in metabolomics

DOI:
[10.1002/cem.3177](https://doi.org/10.1002/cem.3177)

Document Version
Accepted author manuscript

[Link to publication record in Manchester Research Explorer](#)

Citation for published version (APA):
Fu, G., Yi, L., & Pan, J. (2019). LASSObased falsepositive selection for classimbalanced data in metabolomics. *Journal of Chemometrics*. <https://doi.org/10.1002/cem.3177>

Published in:
Journal of Chemometrics

Citing this paper
Please note that where the full-text provided on Manchester Research Explorer is the Author Accepted Manuscript or Proof version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version.

General rights
Copyright and moral rights for the publications made accessible in the Research Explorer are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Takedown policy
If you believe that this document breaches copyright please refer to the University of Manchester's Takedown Procedures [<http://man.ac.uk/04Y6Bo>] or contact uml.scholarlycommunications@manchester.ac.uk providing relevant details, so we can investigate your claim.

LASSO-based false positive selection for class-imbalanced data in metabolomics

Guang-Hui Fu¹, Lun-Zhao Yi², Jianxin Pan^{3*}

¹ School of Science, Kunming University of Science and Technology, Kunming, 650500, P. R. China

² Faculty of Agriculture and Food, Kunming University of Science and Technology, Kunming, Yunnan, 650500, China

³ School of Mathematics, The University of Manchester, Manchester, M13 9PL, UK

Abstract.

Feature selection and re-balancing can be seen as two preprocessing ways in class-imbalanced learning. Recently, There have been many research achievements and applications on LASSO-type feature selection whereas most of them are not directly designed for addressing class-imbalanced data. In this study, we proposed a LASSO-based stable feature selection algorithm for class-imbalanced data analysis, and false positive selection (FPS) under balanced and imbalanced situations was calculated via selection frequency of each predictor in doing stable selection. The results on simulation studies and real data examples show that class-imbalance contributes to avoid over-selection caused by LASSO when the data is highly correlated and a lower FPS can be obtained with class-imbalanced data than balanced one in most of cases in the same settings. A statistical explanation was given for this phenomenon. In addition, it does not need to re-balance the class-imbalanced data for performing such LASSO-based feature selection with a stable strategy, and to some degree, intentionally disequilibrating the balanced data could be an alternative strategy to weaken over-selection and to perform biomarker identification for finding a few of most important biomarkers.

Keywords: class-imbalance; LASSO-based feature selection; re-balance; false positive selection

*Correspondence to: J. Pan. Email: jianxin.pan@manchester.ac.uk

1. Introduction

Feature selection can be utilized as the preprocessing of selecting a subset of candidate predictors and is gaining popularity in class-imbalanced learning in this "big data" era [1–4]. Due to the inherent complex characteristics of class-imbalanced data, performing feature selection from such data requires new understandings and principles to transform vast amounts of raw data efficiently into information and knowledge representation [5]. Feature selection is a critical step due to the high-dimensionality of data across many scientific discoveries, such as QSAR study [6–8] and spectroscopic analysis [9–13]. There are three categories of feature selection in the context of classification, depending on how these feature selection searches combine with the construction of the classification model: filtering [14, 15], wrapping [16, 17] and embedding [1, 2, 4]. The filtering method selects high-ranking features based on statistical or information measures, which is independent of the classifier, whereas it ignores the dependencies among features. The wrapping method wraps a search algorithm around the classification model to search the space of all feature subsets. However wrapping methods are generally computationally intensive as the number of subsets from the feature space grows exponentially as the number of features increases. The embedding method screens out key features while considering the construction of a classifier. Namely, it is integrated in the modelling process and is classifier-dependent [18].

However, there are a few reported works on feature selection methods technically designed to handle the problem of imbalanced class distribution, see references [3, 19–21]. Least absolute shrinkage and selection operator (LASSO) [22] is very popular for variable selection and is widely used for conducting complex data, but just a few of works are directly done for dealing with class-imbalanced data [23, 24].

Another preprocessing strategy for addressing class-imbalanced data is to re-balance the class distribution by sampling [25]. This re-balance can be done by either utilizing under-sampling or over-sampling to equilibrate the sample size of two classes in training data, such as synthetic minority over-sampling technique (SMOTE) [26]. Re-balance techniques with sampling do indeed aid in improving classification accuracy for most imbalanced data sets, but there is still no evidence that re-balancing distribution among classes certainly improve a classifier performance. On the contrary, studies have also shown that classifiers induced from certain imbalanced data sets are comparable to classi-

fiers induced from the same data set balanced by sampling techniques. Here natural questions arise: which preprocessing should be placed on first? Re-balance or LASSO variable selection? is it necessary to re-balance the imbalanced data when doing LASSO feature selection? This questions are important not only for theoretical research, but also for many practical applications. In this paper, we focus on this study by designing a LASSO-based stable feature selection procedure coupled with FPS on simulation studies and real data examples. In the following section, we firstly give the sparse feature selection methods with LASSO-type penalty.

2. LASSO-based sparse regularization logistic regression

Let $\mathbf{y} = (y_1, y_2, \dots, y_n)^T$ be the categorical variable and $\mathbf{X}$ be $n \times (1 + p)$ predictor matrix with the sample size n and the number of predictors p . $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \dots, \beta_p)^T$ is the vector of estimate coefficients. Without loss of generality, assuming $y_i = 1$ if it belongs to minority class (positive) and $y_i = 0$ majority class (negative). $\mathbf{x}_i = (1, x_{i1}, x_{i2}, \dots, x_{ip})^T$ where x_{ij} is the observation value of the i^{th} instance with j^{th} covariate ($i = 1, 2, \dots, n; j = 1, 2, \dots, p$). In linear logistic regression model, the two class-conditional probabilities can be represented as follows:

$$P(y_i = 1|\mathbf{x}_i) = \frac{1}{1 + e^{-\boldsymbol{\beta}^T \mathbf{x}_i}}; \quad P(y_i = 0|\mathbf{x}_i) = \frac{1}{1 + e^{\boldsymbol{\beta}^T \mathbf{x}_i}} \quad (1)$$

Thus,

$$\ln \left(\frac{P(y_i = 1|\mathbf{x}_i)}{P(y_i = 0|\mathbf{x}_i)} \right) = \boldsymbol{\beta}^T \mathbf{x}_i \quad (2)$$

The unknown coefficients $\boldsymbol{\beta}$ are fitted using maximum likelihood estimate, meaningly it needs to maximize the log-likelihood function:

$$\hat{\boldsymbol{\beta}} = \arg \max_{\boldsymbol{\beta}} \left\{ \frac{1}{n} \sum_{i=1}^n (y_i \boldsymbol{\beta}^T \mathbf{x}_i - \ln(\boldsymbol{\beta}^T \mathbf{x}_i)) \right\} \quad (3)$$

Formula (3) can be coupled with LASSO-based penalty (denoted by $C_\alpha(\boldsymbol{\beta})$) to achieve sparse variable selection, and it changes to be

$$\hat{\boldsymbol{\beta}} = \arg \max_{\boldsymbol{\beta}} \left\{ \frac{1}{n} \sum_{i=1}^n (y_i \boldsymbol{\beta}^T \mathbf{x}_i - \ln(\boldsymbol{\beta}^T \mathbf{x}_i)) - \lambda C_\alpha(\boldsymbol{\beta}) \right\}, \quad (4)$$

where $\lambda \geq 0$ is a sparse regularization parameter. Generally, Formula (4) is called sparse regularization logistic regression (SRLR). A widely used form of LASSO-type penalty $C_\alpha(\beta)$ is the following form:

$$C_\alpha(\beta) = \frac{1}{2}(1 - \alpha)\|\beta\|_2^2 + \alpha\|\beta\|_1, \quad (5)$$

where the parameter $\alpha \in [0, 1]$. $\|\beta\|_1 = \sum_{j=1}^p |\beta_j|$ is LASSO penalty ($\alpha = 1$) and $\|\beta\|_2^2 = \sum_{j=1}^p \beta_j^2$ is ridge penalty ($\alpha = 0$). Formula (5) is called elastic-net [27] penalty which is the trade-off between LASSO and ridge regression and can be seen as an improved version of LASSO penalty that is particularly effective for feature selection, especially when the predictors are strongly correlated.

3. Experiment design and Algorithm

3.1 Motivation and experiment design

The model (4) can be solved by cyclical coordinate descent algorithm [28] and the coefficient path of each $\hat{\beta}_j$ is the function of sparse regularization parameter λ (see Fig. 1 (A)). However, an intractable question of this algorithm is that it is typically hard to determine the right amount of parameter λ and the optimal value of λ might be easily changed due to the tiny variation from a certain individual. In other words, the selected features may change significantly due to the influence of a single observation. As such, choosing the right parameter λ is extremely hard for the original path (shown in Fig. 1 (A)). On the contrary, subsampling based model selection is simple and it works in achieving stable feature selection. It was pointed out that the marriage of subsampling and sparse selection algorithms yields finite sample family-wise error control and significantly improves selection methods [29].

The idea of such stable selection strategy is: sparse selection with a subset (randomly chosen from the original data set) is performed many times (T). Each predictor is either selected or not each time, thus the selection frequency (see Fig. 1 (B)) of each predictor can be counted throughout T times replications. Finding the right regularization parameter is replaced by asking how often at least one predictor is selected in iterations. Predictors with maximal selection frequencies are most important.

Therefore, subsampling coupled with sparse regularization is able to replace the regularization parameter λ by selection frequency. The analysis of the selected features in each iteration provides useful information on the importance of the features by using the feature selection frequencies, and the regulariza-

Figure 1: A: Sparse regularization coefficient estimates as a function of λ . The vertical dashed line, denoting the final model selected, needs to be tuned very cautiously. B: Selection frequency of each feature obtained by performing sparse feature selection (A) with subsampling throughout T runs.

tion parameter now is not so important due to the use of selection frequency. Such stable selection strategy make us have a better chance of selecting truly relevant predictors, so subsampling with SRLR is employed to build the algorithm.

3.2 Algorithm

In the settings of high-dimensionality, the estimate coefficient vector β is generally sparse in the sense that only $q < p$ components are non-zero. Namely, most of predictors are uninformative features for a certain task. Here and after that, we call it false positive selection (FPS) if an uninformative predictor is selected as key variable (i.e., its estimate coefficient is non-zero). This study considers FPS on both balanced and imbalanced data. In other words, the amount of uninformative features which are placed in the first through their selection frequency are calculated. The computation is described in algorithm 1 where

n_0 and n_1 respectively are the size of the majority and the minority, and the imbalance ratio $IR = n_1 : n_0$.

Let $f_1 \geq f_2 \geq \dots \geq f_q$ be the selection frequencies of q key features, and z_j be the selection frequency of j^{th} uninformative feature ($j = 1, 2, \dots, g$ and $q + g = p$). We define

$$u_k = \begin{cases} |\{z_j | z_j > f_k\}|, & k = 1, \\ |\{z_j | z_j \geq f_k\}|, & k = 2, \dots, q, \end{cases} \quad (6)$$

where $|\cdot|$ is the cardinal number of the set. Thus, u_k substantially assesses the number of uninformative features with equal to or larger selection frequencies than that of k out of q key features after ranking decreasingly. In this study, $q = 7$ and u_k with $k \in \{1, 2, 4, 6, 7\}$ is counted. Here an example is given for calculating u_k : Assuming that $f_1 = 0.99$, $f_2 = 0.96$, $f_3 = 0.96$, $f_4 = 0.95$, $f_5 = 0.94$, $f_6 = 0.93$, $f_7 = 0.92$, and the maximal 4 selection frequencies of irrelevant variables are $z_1 = 0.96$, $z_2 = 0.94$, $z_3 = 0.93$ and $z_4 = 0.91$, respectively, then

$$\begin{aligned} u_1 &= |\{z_j | z_j > f_1\}| = |\emptyset| = 0; \\ u_2 &= |\{z_j | z_j \geq f_2\}| = |\{z_1\}| = 1; \\ u_4 &= |\{z_j | z_j \geq f_4\}| = |\{z_1\}| = 1; \\ u_6 &= |\{z_j | z_j \geq f_6\}| = |\{z_1, z_2\}| = 2; \\ u_7 &= |\{z_j | z_j \geq f_7\}| = |\{z_1, z_2, z_3\}| = 3. \end{aligned}$$

Actually, u_1 is the amount of uninformative predictors with the selection frequency that is strictly larger than the largest selection frequency of key features. u_k ($k = 2, 4, 6, 7$) is the number of uninformative predictors with the selection frequency that is equal to or larger than that of top k key features. u_q is to measure how many uninformative features will, at least, be involved in order to make sure that all key features are selected. Algorithm 1 contains both inner and outer loops. The inner loop performs stable LASSO feature selection with sparse regularization, and the outer loop is to repeat such stable selection procedure T times. $T = 30$ in this study. It should be pointed out that we fix the negative samples in performing the inner loop in order to effectively concentrate on exploring the effect from the class-imbalance. The imbalance ratio is controlled by choosing the number of negative and positive classes. 2/3 of two class samples are used in both outer and inner loops, so the imbalance ratio keeps invariant in the whole flow. We measure and test the concern with

this procedure over simulation study and real data examples. Simulation study is firstly performed in the following section.

Algorithm 1 Computation of FPS for class-imbalanced data

- 1: parameters n_0 , IR and T ;
 - 2: initial $\mathbf{u} \leftarrow (0, 0, 0, 0, 0)^T$;
 - 3: **for** $t = 1$ to T **do**
 - 4: generate randomly the $\mathbf{X}$ -matrix and label $\mathbf{y}$, then randomly picking out n_0 sample with $y = 0$ and $IR \times n_0$ sample with $y = 1$.
 - 5: initial $\mathbf{w} \leftarrow (0, 0, \dots, 0)^T \in R^p$;
 - 6: Randomly select $2n_0/3$ negative samples;
 - 7: **for** each $i \in [1, 100]$ **do**
 - 8: Randomly select $IR \times 2n_0/3$ positive samples;
 - 9: Do feature selection with SRLR. $w_j \leftarrow w_j + 1$ if $\hat{\beta}_j \neq 0$ for all $j \in \{1, 2, \dots, p\}$, where $\hat{\beta}_j$ is the SRLR estimate coefficient of the j^{th} predictor;
 - 10: **end for**
 - 11: count the number of uninformative features whose selection frequencies are equal to or larger than that of 1, 2, 4, 6, q out of q key features after ranking decreasingly and denoted it by $\Delta\mathbf{u}$. Then $\mathbf{u} \leftarrow \mathbf{u} + \Delta\mathbf{u}$.
 - 12: **end for**
 - 13: $\mathbf{u} \leftarrow \mathbf{u}/T$.
-

4. Simulation study

In this section, we test model selection of the proposed LASSO-based feature selection method coupled with simulation data under both balanced and imbalanced situations. With the simulation data, we actually know which variables are truly important or not.

4.1 Simulation data

In this study, the $\mathbf{X}$ -matrix is simulated by using multivariate normal distribution, namely, $\mathbf{X} \sim N_p(\mathbf{0}, \Sigma)$. For logistic regression, the $\mathbf{y}$ -label is 0-1 vector and is generated via the Bernoulli distribution with probability $p = P(y_i = 1|\mathbf{x}_i) = \frac{1}{1+e^{-\beta^T \mathbf{x}_i}}$, where β is a parameter vector whose component $\beta_j \neq 0$ if j^{th} predictor is a key feature, otherwise $\beta_j = 0$ for all $j = 1, 2, \dots, p$. 7 key features are also utilized and their corresponding $\beta_j = 3, 3, 2.7, 2.7, 2.4, 2.4, 2.1$, respectively. Noted that here a big total sample size

Figure 2: Selection frequency on six cases. A: $n_0 = 240|p = 50|\rho = 0$; B: $n_0 = 240|p = 50|\rho = 0.6$; C: $n_0 = 960|p = 50|\rho = 0.9$; D: $n_0 = 240|p = 400|\rho = 0.3$; E: $n_0 = 960|p = 400|\rho = 0.7$; F: $n_0 = 240|p = 400|\rho = 0.5$.

Table 1: FPS with simulation data. The results under different correlation coefficients are calculated based on 30 replication.

$n_0 \& p$	ρ	IR	u_1	u_2	u_4	u_6	u_7	$n_0 \& p$	ρ	IR	u_1	u_2	u_4	u_6	u_7
$n_0 = 240$ $p = 50$	0	1 : 1	0.00	0.00	0.00	0.00	0.00	$n_0 = 240$ $p = 400$	0	1 : 1	0.00	0.10	0.10	0.10	0.10
		1 : 2	0.00	0.00	0.00	0.00	0.00			1 : 2	0.00	0.00	0.00	0.00	0.00
		1 : 4	0.00	0.00	0.00	0.00	0.00			1 : 4	0.00	0.00	0.00	0.00	0.00
		1 : 8	0.00	0.00	0.00	0.00	0.00			1 : 8	0.00	0.00	0.00	0.00	0.00
	0.2	1 : 1	0.00	0.23	0.23	0.23	0.27		0.2	1 : 1	0.00	0.15	0.15	0.15	0.25
		1 : 2	0.00	0.10	0.10	0.10	0.17			1 : 2	0.00	0.00	0.00	0.00	0.00
		1 : 4	0.00	0.00	0.00	0.00	0.07			1 : 4	0.00	0.00	0.00	0.00	0.05
		1 : 8	0.00	0.00	0.00	0.03	0.07			1 : 8	0.00	0.00	0.00	0.10	0.40
	0.5	1 : 1	0.00	0.27	0.27	0.30	0.47		0.5	1 : 1	0.00	0.25	0.30	0.75	6.90
		1 : 2	0.00	0.00	0.00	0.03	0.20			1 : 2	0.00	0.00	0.00	0.60	9.55
		1 : 4	0.00	0.00	0.00	0.03	0.33			1 : 4	0.00	0.00	0.15	1.05	10.70
		1 : 8	0.00	0.00	0.00	0.20	0.50			1 : 8	0.00	0.00	0.35	2.60	33.55
	0.7	1 : 1	0.00	0.23	0.27	0.53	2.93		0.7	1 : 1	0.05	0.35	1.70	10.30	42.60
		1 : 2	0.00	0.03	0.03	0.47	3.13			1 : 2	0.10	0.30	1.80	14.75	56.55
		1 : 4	0.00	0.00	0.07	0.73	4.37			1 : 4	0.15	0.25	2.35	21.55	65.95
		1 : 8	0.00	0.03	0.23	0.90	5.37			1 : 8	0.10	0.50	3.50	28.65	91.00
	0.9	1 : 1	0.03	0.47	2.07	7.97	15.80		0.9	1 : 1	0.35	2.80	31.40	168.95	270.90
		1 : 2	0.00	0.33	1.73	7.67	15.63			1 : 2	0.40	2.30	40.00	126.95	237.35
		1 : 4	0.07	0.37	2.27	9.33	16.53			1 : 4	0.15	2.75	43.30	116.60	253.60
		1 : 8	0.03	0.37	2.60	10.83	18.20			1 : 8	0.55	5.65	52.25	141.75	237.80

is needed in order to obtain enough negative and positive samples used in our algorithm. The R code for generating this simulation data can be found in the supplementary information.

4.2 Computation and results

SRLR combined with the above simulation data is used to implement the LASSO stable feature selection and assess the FPS. Receiver operating characteristics (ROC) and its area under the curve (AUC) [30–32] are used as measurement to determine the optimal λ in model (4) when doing sparse regularization regression with each subsampling dataset. The influence of class-imbalance for feature selection is comprehensively measured by changing three indices: the sample size, dimensionality of predictors as well as the complexity of covariates. The negative sample size $n_0 = 240, 360, 480, 720, 960$, respectively, the total number of predictors $p = 50$ and 400 , and the correlation coefficients $\rho \in \{0, 0.2, 0.5, 0.7, 0.9\}$. Selection frequency is counted to perform the FPS: which and how many uninformative features are incorrectly identified as key features. The number of key features keeps 7 in all the situations, and the positive to negative ratio here sets be 1 : 1, 1 : 2, 1 : 4 and 1 : 8.

Table 1 lists the results on FPS based on SRLR coupled with the simulation

Table 2: FPS with simulation data. The results under different sample size are calculated based on 30 replication.

$\rho \& p$	n_0	IR	u_1	u_2	u_4	u_6	u_7	$\rho \& p$	n_0	IR	u_1	u_2	u_4	u_6	u_7
$\rho = 0.9$ $p = 50$	360	1 : 1	0.00	0.70	1.10	4.23	11.23	$\rho = 0.7$ $p = 400$	360	1 : 1	0.00	0.55	0.95	3.35	14.05
		1 : 2	0.10	0.23	0.70	4.20	10.93			1 : 2	0.00	0.00	0.35	4.40	23.95
		1 : 4	0.07	0.20	0.73	4.23	11.40			1 : 4	0.00	0.00	0.45	7.65	24.50
		1 : 8	0.13	0.23	1.33	4.87	14.23			1 : 8	0.00	0.10	1.55	15.10	39.15
	480	1 : 1	0.00	0.57	0.70	1.63	3.83		480	1 : 1	0.00	0.40	0.55	0.70	2.70
		1 : 2	0.03	0.07	0.23	1.13	3.67			1 : 2	0.00	0.00	0.05	0.30	4.10
		1 : 4	0.03	0.03	0.20	1.83	3.70			1 : 4	0.00	0.00	0.00	0.45	6.30
		1 : 8	0.03	0.10	0.53	1.90	5.67			1 : 8	0.00	0.00	0.10	0.90	14.85
	720	1 : 1	0.00	0.57	0.57	0.97	2.17		720	1 : 1	0.00	0.25	0.25	0.30	0.40
		1 : 2	0.00	0.07	0.10	0.57	1.77			1 : 2	0.00	0.00	0.00	0.00	0.00
		1 : 4	0.00	0.00	0.00	0.63	2.43			1 : 4	0.00	0.00	0.00	0.00	0.05
		1 : 8	0.00	0.00	0.10	0.63	3.77			1 : 8	0.00	0.00	0.00	0.00	0.40
	960	1 : 1	0.00	0.30	0.30	0.53	1.37		960	1 : 1	0.00	0.55	0.55	0.55	0.65
		1 : 2	0.00	0.03	0.03	0.07	0.80			1 : 2	0.00	0.00	0.00	0.00	0.00
		1 : 4	0.00	0.00	0.00	0.13	1.13			1 : 4	0.00	0.00	0.00	0.00	0.00
		1 : 8	0.00	0.00	0.00	0.43	1.70			1 : 8	0.00	0.00	0.00	0.00	0.00

Table 3: FPS on three real datasets. The results under different class-imbalance ratios are calculated based on 30 replication.

Dataset	IR	u_1	u_2	u_4	u_6	u_7	selected features	reference features
NPC	40:80	0.00	1.00	1.60	4.40	8.90	14, 20, 7, 22, 21, 11, 5	14, 18, 20, 21, 8, 5, 22
	20:80	0.00	0.27	0.97	4.10	6.17	14, 20, 22, 21, 7, 11, 18	
	10:80	0.00	0.03	0.93	3.17	4.87	14, 20, 7, 22, 21, 18, 8	
TBI	27:54	0.00	0.53	0.97	32.37	--	32, 23, 29, 38, 18, 21	29, 23, 20, 38, 32, 36
	16:54	0.00	0.17	1.20	27.27	--	32, 23, 29, 18, 38, 21	
	9:54	0.03	0.10	2.00	19.77	--	29, 32, 23, 18, 21, 38	
TLS	12:24	0.00	0.07	0.13	2.40	--	126, 142, 141, 182, 138, 32	142, 138, 182, 126, 141, 33
	6:24	0.00	0.00	2.03	12.63	--	138, 142, 126, 139, 141, 105	
	3:24	0.00	0.00	17.60	178.17	--	142, 138, 139, 106, 55, 46	

data. It can be seen that the increased number of uninformative features are selected with a rise in ρ , no matter whether the data is class balanced or not, and u_1 is generally equal to zero when the correlation coefficient is not too high, namely the feature with the maximal selection frequency is a key feature. However, if coming to the comparison of balance with imbalance, a lower FPS is obtained with class-imbalanced data than balanced one in most of cases, especially when $IR = 1 : 2$ and $1 : 4$. In addition, u_2 is much smaller in class-imbalanced than balanced cases in the same settings. It means that class-imbalance contributes to avoid over-selection on uninformative features by LASSO in such settings and it will be beneficial for biomarker identification if it just concentrates on which the most important one or few features are.

Table 2 reports the results on false positive selection with different sample size. It can be seen that the increase of sample size can reduce FPS, and u_1 is almost zero with the increase of sample size under both balanced and imbalanced cases. In the same setting, u_2 and u_4 are lower by imbalanced data compared with the balanced one.

4.3 Discussion

Generally, FPS can be reduced by increase of sample size, decrease of the dimensionality of predictors as well as decline of correlation among predictors. Fig. 2 shows the selection frequencies on six scenarios by changing the above three indices. Fig. 2 (A) is a desirable case where all the key features are placed on the top under both balanced and imbalanced conditions due to the low dimensionality and no correlations among predictors. In contrast, Fig. 2 (F) stands for the results from small samples, high dimensionality as well as strongly correlated structure, where just 2, 3, 2, 2 key features place in the front of all irrelevant predictors in a balanced condition and three imbalanced cases, respectively. In Fig. 2 (B) to (E), even though all the key features are still on the top, one irrelevant predictor is also intertwined with them in the balanced condition, and class-imbalance can weaken this mixture and the key features are still on the top.

Therefore, we conclude from these simulation studies that LASSO-based feature selection method tends to be "mild" when performing the balanced data, namely, while more key features are more likely selected, the irrelevant predictors are more likely selected, too, especially for the strongly correlated data. In the imbalanced situation, it becomes "rigorous", in other words, both key features and uninformative predictors are less likely selected, but the real key features will still stand out in stable selection due to their importance to the task. To sum up, class-imbalance probably contributes to avoid over-selection caused by LASSO and it would be beneficial for LASSO-based biomarker identification where the target is to find the most important one or few features.

5. Real data study

5.1 NPC

Nasopharyngeal carcinoma (NPC) dataset [33] contains 100 patients with nasopharyngeal carcinoma and 100 healthy adults in the health control group. The samples from both healthy donors and patients were collected by clinical nurses from the Xiangya Hospital of Hunan, Changsha city, China. 24

metabolites, which were involved in the metabolic processes related to sugars, organic acids, amino acids, fatty acids and sterols, were measured by gas chromatography-mass spectrometry (GC-MS) in all the 200 subjects (See SI-table 1 in supplementary information). NPC is an exactly balanced dataset. Actually, the number of patient samples is exactly equal to the number of healthy samples. In the reference [33], the top seven most important features are No.14, 18, 20, 21, 8, 5, 22 respectively.

In order to test the performance of feature selection under class-imbalance, the above seven biomarkers are used as key features, and the NPC dataset is intentionally disequilibrated. To be specific, off all the 200 samples, 80 out of 100 control samples and 40 (or 20, 10) out of 100 patients are used in each scenario, thus, the positive to negative ratio is 1 : 2, 1 : 4, 1 : 8, respectively. The selected features and FPS are shown in Table 3. We can see that most of the selected key features are consistent with the results obtained via the exactly balanced data, even the class-imbalance ratio changes from 1:2 to 1:8. In three imbalanced cases, 5 of the 7 top selected features are same as selected by utilizing balanced data. Therefore, SRLR with subsampling is suitable for finding key features even if the data is class-imbalanced. On the other hand, u_1 equals zero in three imbalanced cases, it indicates that SRLR with subsampling can still find the most important feature in these imbalanced situations. u_2 , u_4 , u_6 and u_7 decrease as the class-imbalance intensifies, so LASSO-based feature selection algorithm tends to be "rigorous" in serious class-imbalance. Namely, both important variables and irrelevant predictors become hard to be selected in imbalanced case.

5.2 TBI

Traumatic brain injury (TBI) dataset is cited from the literature [34]. The dataset included 73 TBI patients with cognitive impairment (TBI-CI). In addition, 67 healthy controls were included in the investigation. A total of 42 metabolites, which were involved in the metabolic processes related to amino acids, carbohydrates, energy, lipids, organic acids, and urea, were measured by GC-MS in all the 140 subjects (See SI-table 2 in supplementary information). The correct identification of TBI patients with cognitive impairment and determination of the candidates of biomarkers were extremely important in this case. TBI is considered an approximately balanced dataset (67 vs. 73). Six metabolites (No. 29, 23, 20, 38, 32, 36) were identified as novel biomarkers. We also tests the performance of feature selection by artificially disequilibrating the data. Namely, off all the 140 samples, 54 out of the 67 control samples

and 27 (or 16, 9) out of the 73 patients were used in each scenario, respectively. Thus the positive to negative ratio is 1 : 2, 8 : 27 and 1 : 6. The selected features in these situations are shown in Table 3, and we found that the key features obtained from the balanced case can be almost selected in class-imbalanced conditions. In three imbalanced conditions, 4 of the 6 top selected features are same as selected by using balanced data. Thus most of the relevant variables can be selected in imbalanced case by SRLR coupled with subsampling. u_1 is almost zero and u_2 is very small, this means that one or two most important biomarkers can be almost exactly discovered under class-imbalanced situations.

5.3 TLS

Tubulointerstitial Lesions' Severity (TLS) data is obtained by Department of Nephrology of the University Hospital of Ioannina, Greece [35]. The data is measured by GC-MS analysis on the urinary samples collected from 25 patients with mild tubulointerstitial lesions and 25 patients with severe tubulointerstitial lesions. There are 200 variables, but most of them are uninformative or redundancy. According to the work in [36], the first top six key are No. 142, 138, 182, 126, 141, 33, respectively.

This dataset is exactly balanced but is still disequibrated artificially, to be specific, off all the 50 subjects, 24 out of 25 mild tubulointerstitial lesions patients as well as 12 (or 6, 3) out of 25 severe tubulointerstitial lesions patients are drawn in terms of our experiment design, and the positive to negative ratio is 1 : 2, 1 : 4 and 1 : 8, respectively. FPS and the selected features in these situations are shown in Table 3. It can be seen that u_1 is zero and u_2 is almost zero at three class-imbalanced cases. In other words, the variables with top one or two selection frequencies are really relevant features. Off all the 6 top selected features, 5, 4 and 2 key variables are selected in corresponding three imbalanced cases. In addition, the most important variables (such as No. 142) are almost still selected with large selection frequency. In the last case where the size of positive class is just 3 and the imbalance is extremely serious, whereas two most important predictors are still picked out from 200 candidate predictors.

The results from all three real data sets show that SRLR with subsampling is able to find important variables in class-imbalanced scenarios. In addition, the irrelevant variables is extremely hard to rank top even if in "small sample, serious imbalance" case. Therefore SRLR with subsampling is suitable for identifying those most important biomarkers in the class-imbalanced settings.

6. Discussion and conclusion

6.1 Statistical explanation

Class-imbalance contributes to avoid over-selection caused by LASSO when the data is highly correlated. This might be the following reasons: Firstly, disequilibrating the balanced data results in the increase of noises, and this variation would be beneficial to avoid over-selection and, as a consequence, might reduce the false positive selection. This phenomenon is very similar with the case from neural networks where such strategy of increase of noise is employed to restrain over-fitting [37]. Another reason might be the marriage of subsampling with LASSO, this is termed as stability selection in the reference [29]. This stable selection can put a large emphasis on avoiding false positive selection. In fact, for a very extreme case, such as the positive to negative ratio reaches to 1: 100, it is still possible to find out those most important features. For example, we want to identify the biomarkers of coronary heart disease according to just 1 patient and 100 healthy controls. Obviously, the patient is very critical, but it is unlikely that the measurement from this patient nicely stands for the mean of many patients suffered from this disease. In other words, compared with the measurement from controls, the magnitude of some predictors from this patient is extended and consequently it is more likely to be selected, and some is reduced and it gets easy to be eliminated. However, the selection frequencies of all these variable are decreased when doing replications and finally they are all less likely to be selected. On the other hand, Although their selection frequencies are declined as a whole, the frequency of a biomarker is still larger than that of irrelevant predictors, so a biomarker is still probably placed top and is firstly selected. Obviously, performing selection from such skewed data makes us at the risk of selecting uninformative predictor as key feature due to the variation from this extreme class-imbalance, however, this risk can be avoided to a great extent due to the application of stable selection.

6.2 Conclusion

It has been shown that doing LASSO-based stable feature selection in class-imbalance learning is not necessary to re-balance the majority and the minority in advance. The performance in class-imbalanced condition measured by FPS was ever better than that in a 1 : 1 class ratio situation. Biomarker identification in finding a few of most important biomarkers could also be done in class-

imbalanced situation. In addition, disequilibrating the class ratio might be an alternative to reduce over-selection where the data are strongly correlated. We consider that these findings are important not only for theoretical research development, but also for many practical application scenarios. In this study, the variable selection method is limited to the LASSO-type sparse selection, how about other types of variable selection procedures? is it necessary or not to re-balance the class-imbalanced data with general feature selection algorithms? this is our next research study.

• Acknowledgements

We sincerely thank the editor and two anonymous reviewers for their constructive comments that led to the current improved version of the paper. This work is financially supported by the National Natural Science Foundation of China (Grant Nos. 11761041, 21465016, and 21775058). The studies meet with the approval of the university's review board.

References

- [1] H. Mamitsuka. Selecting features in microarray classification using roc curves. *Pattern Recognition*, 39(12):2393–2404, 2006.
- [2] Z. Liu and M. Tan. Roc-based utility function maximization for feature selection and classification with applications to high-dimensional prostate data. *Biometrics*, 64(4):1155–1161, 2008.
- [3] L. Yin, Y. Ge, K. Xiao, X. Wang, and X. Quan. Feature selection for high-dimensional imbalanced data. *Neurocomputing*, 105:3–11, 2013.
- [4] P. Zhou, X. Hu, P. Li, and X. Wu. Online feature selection for high-dimensional class-imbalanced data. *Knowledge-Based Systems*, 136:187–199, 2017.
- [5] H. He and E. A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009.
- [6] M. Shahlai. Descriptor selection methods in quantitative structure–

- activity relationship studies: a review study. *Chemical reviews*, 113(10):8093–8103, 2013.
- [7] A. V. Zakharov, M. L. Peach, M. Sitzmann, and M. C. Nicklaus. Qsar modeling of imbalanced high-throughput screening data in pubchem. *J. Chem. Inf. Model.*, 54(3):705–712, March 2014.
 - [8] Y. Wang, J.-J. Huang, N. Zhou, D.-S. Cao, J. Dong, and H.-X. Li. Incorporating pls model information into particle swarm optimization for descriptor selection in qsar/qspr. *Journal of Chemometrics*, 29(12):627–636, 2015.
 - [9] G.H. Fu, Q.S. Xu, H.D. Li, D.S. Cao, and Y.Z. Liang. Elastic net grouping variable selection combined with partial least squares regression (en-plsr) for the analysis of strongly multi-collinear spectroscopic data. *Applied Spectroscopy*, 65(4):402–408, 2011.
 - [10] E. Andries and S. Martin. Sparse methods in spectroscopy: An introduction, overview, and perspective. *Applied spectroscopy*, 67(6):579–589, 2013.
 - [11] R. M. Balabin and S. V. Smirnov. Variable selection in near-infrared spectroscopy: Benchmarking of feature selection methods on biodiesel data. *Analytica chimica acta*, 692(1):63–72, 2011.
 - [12] X. Zou, J. Zhao, J.W. P. Malcolm, H. Mel, and H. Mao. Variables selection methods in near-infrared spectroscopy. *Analytica Chimica Acta*, 667(1):14 – 32, 2010.
 - [13] Y.-H. Yun, H.-D. Li, B.-C. Deng, and D.-S. Cao. An overview of variable selection methods in multivariate analysis of near-infrared spectra. *TrAC Trends in Analytical Chemistry*, 113:102 – 115, 2019.
 - [14] W. Cai, Y. Li, and X. Shao. A variable selection method based on uninformative variable elimination for multivariate calibration of near-infrared spectra. *Chemometrics and Intelligent Laboratory Systems*, 90(2):188 – 194, 2008.
 - [15] S. Favilla, C. Durante, M. L. Vigni, and M. Cocchi. Assessing feature relevance in npls models by vip. *Chemometrics and Intelligent Laboratory Systems*, 129:76 – 86, 2013. Multiway and Multiset Methods.

- [16] B. T Luke. Evolutionary programming applied to the development of quantitative structure-activity relationships and quantitative structure-property relationships. *Journal of Chemical Information and Computer Sciences*, 34(6):1279–1287, 1994.
- [17] M. Arakawa, Y. Yamashita, and K. Funatsu. Genetic algorithm-based wavelength selection method for spectral calibration. *Journal of chemometrics*, 25(1):10–19, 2011.
- [18] Y. Saeys, I. Inza, and P. Larrañaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, 2007.
- [19] S. Ma and J. Huang. Regularized roc method for disease classification and biomarker selection with microarray data. *Bioinformatics*, 21(24):4356–4362, 2005.
- [20] S. Maldonado, R. Weber, and F. Famili. Feature selection for high-dimensional class-imbalanced data sets using support vector machines. *Information Sciences*, 286:228–246, 2014.
- [21] Z. Zheng, X. Wu, and R. Srihari. Feature selection for text categorization on imbalanced data. *ACM Sigkdd Explorations Newsletter*, 6(1):80–89, 2004.
- [22] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 58(1):267–288, 1996.
- [23] H. Wang, Q. Xu, and L. Zhou. Large unbalanced credit scoring using lasso-logistic regression ensemble. *PLOS ONE*, 10(2):1–20, 02 2015.
- [24] I. Ahmed, A. Pariente, and P. Tubert-Bitter. Class-imbalanced subsampling lasso algorithm for discovering adverse drug reactions. *Statistical Methods in Medical Research*, 27(3):785–797, 2018. PMID: 27114328.
- [25] A. Ali, S.M. Shamsuddin, and A.L. Ralescu. Classification with class imbalance problem: A review. *INTERNATIONAL JOURNAL OF ADVANCES IN SOFT COMPUTING AND ITS APPLICATIONS*, 7(3), 2015.
- [26] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

- [27] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 67(2):301–320, 2005.
- [28] J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332, 2007.
- [29] N. Meinshausen and P. Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.
- [30] T. Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861 – 874, 2006.
- [31] J. A Hanley and B. J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.
- [32] K. Boyd, K. H Eng, and C D. Page. Area under the precision-recall curve: Point estimates and confidence intervals. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 451–466. Springer, 2013.
- [33] L. Yi, C. Song, Z. Hu, L. Yang, L. Xiao, Bin Yi, W. Jiang, Y. Cao, and L. Sun. A metabolic discrimination model for nasopharyngeal carcinoma and its potential role in the therapeutic evaluation of radiotherapy. *Metabolomics*, 10(4):697–708, 2014.
- [34] L. Yi, Y. Shi, S. and Wang, W. Huang, Z. Xia, Z. Xing, and Z. Peng, W. and Wang. Serum metabolic profiling reveals altered metabolic pathways in patients with post-traumatic cognitive impairments. *Scientific reports*, 6, 2016.
- [35] N. G. Psihogios, R. G. Kalaitzidis, S. Dimou, K. I. Seferiadis, K. C. Siamopoulos, and E. T. Bairaktari. Evaluation of tubulointerstitial lesions’ severity in patients with glomerulonephritides: An nmr-based metabonomic study. *Journal of Proteome Research*, 6(9):3760–3770, 2007. PMID: 17705523.
- [36] Y.-H. Yun, B.-C. Deng, D.-S. Cao, W.-T. Wang, and Y.-Z. Liang. Variable importance analysis based on rank aggregation with applications in

metabolomics for biomarker discovery. *Analytica Chimica Acta*, 911:27–34, 2016.

- [37] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, and J. Schmidhuber. A novel connectionist system for unconstrained handwriting recognition. *IEEE transactions on pattern analysis and machine intelligence*, 31(5):855–868, 2009.

Supplementary Information

Supplementary data and R code can be found in the supplementary material.