Chemometrics and Intelligent Laboratory Systems xxx (2017) 1–10

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics

Stable variable selection of class-imbalanced data with precision-recall criterion

Guang-Hui Fu [a], Feng Xu [a], Bing-Yang Zhang [a], Lun-Zhao Yi [b,*]

[a] School of Science, Kunming University of Science and Technology, Kunming, 650500, PR China
[b] Yunnan Food Safety Research Institute, Kunming University of Science and Technology, Kunming, 650500, PR China

ARTICLE INFO

Keywords:
Precision-recall curve
Class-imbalanced data
Sparse regularization logistic regression
Stable variable selection
Subsampling

ABSTRACT

Screening important variables for class-imbalanced data is still a challenging task. In this study, we propose an algorithm for stably selecting key variables on class-imbalanced data based on the precision-recall curve (PRC), where the PRC is utilized as the assessment criterion in the model building stage, and sparse regularized logistic regression combined with subsampling (SRLRS) is designed to perform stable variable selection. Considering the characteristic of class-imbalanced data, we also proposed classification-based partition for cross validation, as well as leaving half of majority observations out and leaving one minority observation out (LHO-LOO) for subsampling. Simulation results and real data showed that our algorithm is highly suitable for handling class-imbalanced data, and that the PRC can be an alternative evaluation criterion for model selection when handling class-imbalanced data.

1. Introduction

Class-imbalanced data arise from many fields of recent scientific discoveries, and common classifiers are widely employed to handle these data. However, these common classifiers usually perform poorly for class-imbalanced data due to the lack of attention to the category of minority [1]. In fact, most of the existing classifiers assume that the underlying training set is evenly distributed, and in practice, the classifier operates on data drawn from the same distribution as the training data [2]. In class-imbalanced classification, the training set for one class (majority) significantly surpasses that of the other class (minority) in which the latter is often more interesting. Using the classifiers produced by standard machine learning algorithms without adjusting for studying class-imbalanced data may be a critical mistake [2].

Owing to the overabundance of negative (majority) examples in class-imbalanced data, classification accuracy is not a sensible evaluation measure because it overvalues the always-negative classifier [3]. Generally, evaluation of classifiers through the receiver operating characteristic (ROC) curve is common in binary classification [4,5]. The ROC curve is a threshold-free measure, and it shows pairs of specificity and sensitivity values calculated at all possible threshold scores [6]. Techniques to develop classifiers that optimize the area under the ROC curve have also been proposed [7–10]. The ROC curve is useful because it provides a visual representation of the relative tradeoffs between the benefits (reflected by true positives) and costs (reflected by false positives) of classification with regard to data distributions. However, for class-imbalanced data, the ROC curve tends to provide an overly optimistic view of the performance of an algorithm [1]. Realizing its disadvantages in dealing with class-imbalanced data, precision-recall curve (PRC) is more informative than ROC, and it has become a basis for assessing classification methods on class-imbalanced data [6]. The PRC has recall on the x-axis and precision on the y-axis. Precision is the proportion of true positives among the positive predictions, and recall measures the proportion of positives that are correctly identified as such. The potential advantages of PRC on class-imbalanced data are due to the fact that the PRC ignores true negatives altogether. As a result, PRC is suitable to assess the performance of the minority samples. As an alternative to ROC, the area under the precision-recall curve (AUPRC) can also be employed as a performance metric. Boyd et al. [11] presented a detailed overview of the methods to compute the area under the empirical precision-recall curve, including lower trapezoid, average precision, interpolated median, and confidence interval estimation. Although discussion of the PRC appears in the literature, its usage has been solely for evaluation in the final model. However, given that PRC

* Corresponding author.
E-mail addresses: guanghuifu@kmust.edu.cn (G.-H. Fu), yilunzhao@kmust.edu.cn (L.-Z. Yi).

https://doi.org/10.1016/j.chemolab.2017.10.015
Received 25 May 2017; Received in revised form 5 September 2017; Accepted 18 October 2017
Available online xxxx

is recommended to evaluate classifiers, using this criterion seems natural in estimating the classifier directly. In this study, the PRC is used as a criterion to screen key variables in the stage of model building.

Another feature of class-imbalanced data is high-dimensionality, which increases the difficulty of statistical modeling for class-imbalanced data. Generally, just a few of all the candidate covariates are important for a certain response, and the rest are redundancies or noises. The mix of such uninformative variables often leads to decreased classification accuracy. Even worse, high-dimensionality leads to class overlapping [12,13]. When overlapping patterns are present in each class for some feature space, determining discriminative rules to separate the classes is extremely difficult. The overlapping feature space may cause the features to lose their intrinsic property, thereby making them redundant or irrelevant to help recognize good decision boundaries between classes [14]. Variable selection is essential in handling high-dimensionality. Variable selection eliminates the uninformative variables and removes overlapping among the predictors. Although numerous classification and variable selection methods are available [15–24], classification and variable selection of class-imbalanced data remains a challenging task, especially in the case of high-dimensionality. Feature selection in class-imbalanced classification is underexplored, and the lack of a systematic approach to feature selection for class-imbalanced and high-dimensional data opens many research possibilities [25,26].

In this study, we focus on variable selection for class-imbalanced data with high-dimensionality. Sparse regularized logistic regression is used to achieve variable screening, and PRC is introduced at the stage of model building as the tool to determine the best regularized parameters in the sparse regularized model. Class-imbalanced simulation data with different sample size, variable dimension, class-imbalanced ratio, and correlation are generated to simulate and evaluate our algorithm. The advantage of using simulation data is that we know the actual key variables that are related to the outcome, that is, we know which variables should be included in the final model under the ideal situation. We also present a comprehensive comparison of PRC and ROC. Based on simulation data, the result shows that a combination of sparse regularization and PRC is suitable to deal with class-imbalanced data. For the real class-imbalanced data, we further introduced subsampling to sparse regularized logistic regression to perform stable variable selection. On the one hand, we assessed our algorithm by comparing situations of imbalanced data with the balanced situation. On the other hand, false discovery rate (FDR) was employed to evaluate our methods under ROC and PRC. The result shows that our PRC-based algorithm is highly suitable for handling class-imbalanced data.

Nowadays, sparse regularization for variable selection is popular in high-dimensional data analysis [27–33]. However, few studies have been conducted on these sparse methods as applied to class-imbalanced data in metabolomics [32–34], especially when the data are severely class-imbalanced. The results of simulation and real data show that sparse regularized logistic regression with PRC is an alternative when dealing with class-imbalanced data.

2. Theory

2.1. Sparse regularized logistic regression

Let $\{\boldsymbol{x}_i, y_i\}$ be the observation, $(i = 1, 2, \cdots, n)$, and the response be binary. Without loss of generality, assuming that $y_i = 1$ or 1, $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^T$ and x_{ij} is the observation of the ith instance with jth covariate, $(j = 1, 2, \cdots, p)$. In linear logistic regression model, the two class-conditional probabilities can be represented as follows:

$$P(y_i = 1|\boldsymbol{x}_i) = \frac{1}{1 + e^{-(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}_i)}} \tag{1}$$

$$P(y_i = 0|\boldsymbol{x}_i) = 1 - P(y_i = 1|\boldsymbol{x}_i) = \frac{1}{1 + e^{(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}_i)}} \tag{2}$$

This means that

$$\ln\left(\frac{P(y_i = 1|\boldsymbol{x}_i)}{P(y_i = 0|\boldsymbol{x}_i)}\right) = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}_i \tag{3}$$

The unknown coefficients β_0 and $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)^T$ are fitted using maximum likelihood estimate, and it needs to maximize the log-likelihood function:

$$\max_{\beta_0, \boldsymbol{\beta}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[y_i (\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}_i) - \ln(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}_i) \right] \right\} \tag{4}$$

Sparse regularized techniques, such as LASSO [27] and elastic net [29], are of great importance for variable selection. Elastic net penalty is the mix of L_1-norm (LASSO) and L_2-norm (ridge regression), which can be defined as

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

where $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$, $\|\boldsymbol{\beta}\|_2^2 = \sum_{j=1}^{p} \beta_j^2$, and $0 \le \alpha \le 1$. When $\alpha = 1$, $C_\alpha(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1$ is the LASSO penalty, which can shrink the small absolute coefficients to exactly zero. Thus, predictors with zero-value coefficients are eliminated, and those with non-zero coefficients are kept as the important variables (biomarkers). This is the reason why LASSO can perform variable selection. However, the performance of LASSO becomes poor when the predictors are highly correlated [29]. When $\alpha = 0$, $C_\alpha(\boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{\beta}\|_2^2$ is the ridge penalty, which is extremely suitable for correlated data, whereas it cannot select key variables directly. Elastic net penalty $(0 < \alpha < 1)$ is the tradeoff between the ridge regression (L_2-norm) and LASSO regression (L_1-norm), and it is particularly useful for variable selection when the predictors are strongly correlated.

Combining formulas (4) and (5), the objective of sparse regularized logistic regression can be expressed as follows:

$$\max_{\beta_0, \boldsymbol{\beta}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[y_i (\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}_i) - \ln(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}_i) \right] - \lambda C_\alpha(\boldsymbol{\beta}) \right\} \tag{6}$$

Where λ is a regularization parameter and $\lambda \ge 0$. The preceding objective can be solved by cyclical coordinate descent algorithm [35].

2.2. Precision-recall curve (PRC)

When the response is binary, the data are divided into positives and negatives. A binary classifier then predicts all the samples as either positive or negative. Thus, it produces four types of outcome: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). The common evaluation criterion in binary classification is the ROC curve [5,36], and the area under it is usually employed to assess the performance of various classifiers. However, the ROC curve tends to provide an overly optimistic result for class-imbalanced data, especially when the class-skew is severe. With its disadvantages in dealing with imbalanced data, PRC has become an alternative for evaluating classification methods on skewed data [6]. For example, the PRC is used in assessing the final model [3,6,10]. The PRC can provide the viewer with an accurate prediction of future classification performance because it evaluates the fraction of true positives among positive predictions. Moreover, the PRC is more informative than the ROC curve when evaluating binary classifiers on class-imbalanced data [6]. The PRC has recall on the x-axis and precision on the y-axis. Recall, which is also known as true positive rate (TPR), is the proportion of correctly classified positives.