Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 122 (2017) 677–684

Procedia
Computer Science

www.elsevier.com/locate/procedia

Information Technology and Quantitative Management (ITQM 2017)

The Study of Credit Scoring Model Based on Group Lasso

Hongmei Chen[a,] * ,Yaoxin Xiang[a]

[a] School of Statistics, Capital University of Economics and Business, Beijing, China， 100070

Abstract

Credit scoring model is one of common tools for commercial banks to manage credit risks. In this paper, we use a public dataset from UCI machine learning repository and construct credit scoring models based on Group Lasso Logistic Regression, where the tuning parameters λ are selected by the Akaike Information Criterion(AIC), Bayesian Information Criterion(BIC) and Cross Validation prediction errors respectively. The experimental results show that the Group Lasso method is better than backward elimination in both interpretability and prediction accuracy.

Peer-review under responsibility of the scientific committee of the 5th International Conference on Information Technology and Quantitative Management, ITQM 2017.

Keywords: Credit Scoring Model, Credit Risks, Group Lasso;

1. Introduction

Lasso(Least Absolute Shrinkage and Selection Operator), proposed by Tibshirani [1996], is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces. It was based on Breiman's[1995] Nonnegative Garrote. In the field of credit risk management, some commercial banks rely on the experience of credit officers to engage in approvals of loans. Credit scoring model as the core of credit risk management， more and more business systems based on data mining and mathematical statistics on a large scale. In fact, credit scoring model is a kind of statistical model which analyzes a large amount of customers' historical data and extracts the laws and features of credit risk, then constructing an appropriate model to evaluate risk for new applicants or existing customers.

With the rapid development of credit card business, credit scoring model will certainly play an important role in the near future. At the same time, with the development of information technology, more and more data are stored in many fields and different industries. Credit scoring model is one of common tools for commercial banks to manage credit risk, some existing scoring models are often unable to effectively screen out dangerous

* Corresponding author.
E-mail address: chenhm@cueb.edu.cn

1877-0509 © 2017 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of the scientific committee of the 5th International Conference on Information Technology and Quantitative Management, ITQM 2017.
10.1016/j.procs.2017.11.423

customers because of credit information asymmetry in credit card business. However, adding too much customers' information will cause biased estimation and instability of credit scoring model because of too many variables, it is of significance to apply the variable selection method to the development of the credit scoring model. In the early development of the credit scoring model, we need to establish many dummy variables as explanatory variables of the model, the group lasso method executes variable selection on group variables, which enables the dummy variables belonging to the same group could be reserved or eliminated in the model.

2. Related studies

Concerning the penalty term, ridge regression proposed by Hoerl et al. [1970] first exploits it, this method can only perform regularization but not variable selection. Lasso may perform bad when levels of a categorical variable are coded as a collection of binary covariates. In this case, it often doesn't make sense to include only a few levels of the covariate, the group lasso method, proposed by Yuan et al. [2006], can ensure that all the variables encoding the categorical covariate are either included or excluded from the model together.

Fang et al. [2014] make an empirical analysis of personal credit evaluation in China based on a consumption credit default data, and find that lasso logistic regression does better in prediction accuracy and seizing key factors that impacting credit risk, compared with full model and stepwise regression. Hu et al. [2015] point out that it needs to set many dummy variables as covariates in personal credit evaluation, and group lasso allow predefined groups of covariates to be selected into or out of a model together, resulting in better performance in interpretability and prediction accuracy. Zhang et al. [2016] make an empirical analysis using generalized semi-parametric additive model while applying group lasso to selecting variables and estimating coefficients, and find that group lasso does better in classification ability and computation efficiency compared with linear logistic regression.

3. Group lasso

Group lasso was introduced by Yuan and Lin (2006), allowing predefined groups of covariates to be selected into or out of a model together, so that all the members of a particular group are either included or not included. It is very useful in many settings, perhaps the most obvious is when levels of a categorical variable are coded as a collection of binary covariates. Under this circumstance, it usually doesn't make sense to include only a few levels of the covariate. The group lasso can ensure that all the variables encoding the categorical covariate are either included or excluded from the model together. The objective function for the group lasso is a natural generalization of the standard lasso objective

$$\min\left\{\left\|y-\sum_{j=1}^{J}X_j\beta_j\right\|_2^2+\lambda\sum_{j=1}^{J}\left\|\beta_j\right\|_{K_j}\right\},\quad \|z\|_{K_j}=\left(z^tK_jz\right)^{1/2} \tag{1}$$

where the design matrix X and covariate vector β have been replaced by a collection of design matrices X_j and covariate vectors β_j, one for each of the J groups. As for variable selection, the selection of tuning parameter λ is very important, and it is usually selected by the Akaike Information Criterion(AIC), Bayesian Information Criterion(BIC) and Cross Validation prediction errors.

$$\mathrm{AIC}(\lambda)=\log\left(\frac{\|\mathrm{Y}-\mathrm{X}\beta(\lambda)\|_2^2}{\mathrm{n}}\right)+2\times\frac{\mathrm{df}}{\mathrm{n}} \tag{2}$$

$$\mathrm{BIC}(\lambda)=\log\left(\frac{\|\mathrm{Y}-\mathrm{X}\beta(\lambda)\|_2^2}{\mathrm{n}}\right)+\log(\mathrm{n})\times\frac{\mathrm{df}}{\mathrm{n}} \tag{3}$$

$$\mathrm{CV}(\lambda)=\frac{\|\mathrm{Y}-\mathrm{X}\beta(\lambda)\|_2^2}{\mathrm{n}(1-\frac{\mathrm{df}}{\mathrm{n}})^2} \tag{4}$$

where df is effective degrees of freedom.