Bioinformatics, 31(24), 2015, 3991–3993
doi: 10.1093/bioinformatics/btv497
Advance Access Publication Date: 26 August 2015
Applications Note

Genome analysis

glmgraph: an R package for variable selection and predictive modeling of structured genomic data

Li Chen[1,2], Han Liu[3], Jean-Pierre A. Kocher[1], Hongzhe Li[4,*] and Jun Chen[1,*]

[1]Division of Biomedical Statistics and Informatics and Center for Individualized Medicine, Mayo Clinic, Rochester, MN 55905,USA, [2]Department of Computer Science, Emory University, Atlanta, GA 30322,USA, [3]Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA and [4]Department of Biostatistics and Epidemiology, University of Pennsylvania, Philadelphia, PA 19104, USA

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on April 14, 2015; revised on July 23, 2015; accepted on August 17, 2015

Abstract

Summary: One central theme of modern high-throughput genomic data analysis is to identify relevant genomic features as well as build up a predictive model based on selected features for various tasks such as personalized medicine. Correlating the large number of 'omics' features with a certain phenotype is particularly challenging due to small sample size (n) and high dimensionality (p). To address this small n, large p problem, various forms of sparse regression models have been proposed by exploiting the sparsity assumption. Among these, network-constrained sparse regression model is of particular interest due to its ability to utilize the prior graph/network structure in the omics data. Despite its potential usefulness for omics data analysis, no efficient R implementation is publicly available. Here we present an R software package 'glmgraph' that implements the graph-constrained regularization for both sparse linear regression and sparse logistic regression. We implement both the L_1 penalty and minimax concave penalty for variable selection and Laplacian penalty for coefficient smoothing. Efficient coordinate descent algorithm is used to solve the optimization problem. We demonstrate the use of the package by applying it to a human microbiome dataset, where phylogeny structure among bacterial taxa is available.

Availability and implementation: 'glmgraph' is implemented in R and C++ Armadillo and publicly available under CRAN.

Contact: chen.jun2@mayo.edu or hongzhe@upenn.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

With the development of high-throughput technology such as next generation sequencing, biology has advanced into the 'Big Data' era. Though generation of these 'omics' data has become increasingly cheaper, high-throughput analysis of the omics data remains challenging due to the massive amounts of omics features. One central goal of genomic studies is to correlate these omics features with a variable of interest such as disease status. Sparse regression models, which jointly analyze the omics features, have gained popularity for omics data analysis due to its good model interpretability, computational efficiency and superior predictive power (Waldron *et al.*, 2011). The omics features are usually related by an underlying graph or network. For example, genes can be grouped into biological pathways and metagenomic sequences are related by a phylogenetic tree. Incorporating the prior network information has been demonstrated to improve both model selection and predictive power (Chen *et al.*, 2013; Tian *et al.*, 2014). The network-constrained sparse regression model proposed by Li and Li (2008), which smoothes the coefficients with respect to the underlying network structure by imposing a Laplacian penalty, provides an efficient way to utilize the prior structure information. However, no R package of the method has been available. Though it can be cast into a Lasso problem by data augmentation approach for a continuous response, it is not computationally efficient and memory-friendly, especially when the dimension p is high, since it converts an $n \times p$ data matrix into an $(n+p) \times p$ augmented data matrix. The data augmentation approach cannot be extended to handle the binary outcome, and network-constrained sparse logistic regression requires specific algorithm. Moreover, no software exists for the combination of Laplacian penalty and non-convex sparsity penalty such as the minimax concave penalty (MCP), though the superior property of the combination has been demonstrated in linear regression setting (Huang *et al.*, 2011). Therefore, we aim to provide an R package with similar interface as the popularly used 'glmnet' package to implement the network-constrained sparse linear and logistic regression with sparsity options of L_1 penalty and MCP. Direct cyclic coordinate descent algorithm is applied to speed up computation. The R package 'glmgraph' will be useful for variable selection and prediction for high-dimensional datasets when there exists graph/network structure among predictors.

2 Methods

2.1 Penalized log-likelihood approach

Suppose we have n independent samples with p predictors. Denote y_i and $\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^T$ the outcome and predictors of the ith sample. The penalized log-likelihood for network-constrained sparse regression is formulated as

$$pl(\boldsymbol{\beta}; \lambda_1, \lambda_2) = \frac{1}{n}\sum_{i=1}^{n}\{-l(\boldsymbol{\beta}; y_i, x_i)\} + p^{sp}_{\lambda_1}(\boldsymbol{\beta}) + p^{sm}_{\lambda_2}(\boldsymbol{\beta}; \mathcal{G}),$$

where $\boldsymbol{\beta}$ are the regression coefficients, and $p^{sp}_{\lambda_1}(.), p^{sm}_{\lambda_2}(.)$ are sparsity and smoothness penalty function respectively to induce sparsity and smoothness of the solution with respect to the underlying graph structure $\mathcal{G}$. λ_1, λ_2 are tuning parameters controlling the degree of regularization. For sparsity penalty, we include both L_1 penalty and MCP, which are defined as

$$p^{L}_{\lambda_1}(\boldsymbol{\beta}) = \lambda_1 \sum_{\mathrm{j}=1}^{p} |\beta_{\mathrm{j}}|, \;\; p^{M}_{\lambda_1}(\boldsymbol{\beta}; \gamma) = \lambda_1 \sum_{\mathrm{j}=1}^{p} \int_0^{|\beta_j|} (1 - \frac{x}{\gamma\lambda_1})_+ d\mathrm{x},$$

where MCP has an extra parameter γ, which can be fixed or tuned. For smoothness penalty, we use the graph Laplacian penalty

$$p^{G}_{\lambda_2}(\boldsymbol{\beta}) = \lambda_2 \boldsymbol{\beta}^{\mathrm{T}} \mathrm{L} \boldsymbol{\beta},$$

where L is the Laplacian matrix defined with respect to the graph $\mathcal{G}$ For unnormalized Laplacian matrix L, $\boldsymbol{\beta}^{\mathrm{T}} L \boldsymbol{\beta} = \sum_{1 \le j \le k \le p} a_{ij}(\beta_j - \beta_k)^2$, which shrinks coefficients to a common value with the shrinking strength determined by the adjacency coefficient a_{ij}. We implement linear and logistic regression to address both continuous and binary outcomes, where the log-likelihood parts (up to additive constants) are

$$l(\boldsymbol{\beta}; \mathrm{y_i}, \mathbf{x}_\mathrm{i}) = \begin{cases} (\mathrm{y_i} - \mathbf{x}_\mathrm{i}^{\mathrm{T}}\boldsymbol{\beta})^2/2 & \text{linear regression} \\ \mathrm{y_i}(\mathbf{x}_\mathrm{i}^{\mathrm{T}}\boldsymbol{\beta}) - \log(1 + \mathrm{e}^{\mathbf{x}_\mathrm{i}^{\mathrm{T}}\boldsymbol{\beta}}) & \text{logistic regression} \end{cases}$$

The maximum penalized log-likelihood estimate is obtained by

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, pl(\boldsymbol{\beta}; \lambda_1, \lambda_2).$$

2.2 Cyclic coordinate descent algorithm

Since the Laplacian penalty is convex in $\boldsymbol{\beta}$, coordinate descent (CD) algorithm developed for sparse regression with convex and non-convex sparsity penalties (Breheny and Huang, 2011; Friedman *et al.*, 2010) can be readily extended to the graph-constrained case. For linear regression, we have closed-form solution for each coordinate update. For logistic regression, we solve a series of graph-constrained sparse linear regression problems at each iteratively reweighed least squares step. The detailed numeric recipes, as well as the computational complexity, are given in the Supplementary Note. To speed up computation, we also employ the 'active set' idea: after a complete cycle through all the variables, we iterate on only the non-zero variables (active set) till convergence. If another complete cycle does not change the active set, we are done, otherwise the process is repeated. For a continuous outcome, direct application of CD algorithm achieves significant computational speed-up over the data augmentation approach using `glmnet` (Supplementary Table S1).

3 Example

To demonstrate the potential performance gain by exploiting the underlying graph structure among predictors using our package, we compare three competing methods `glmgraph` (MCP + Laplacian), `glmnet` (Friedman *et al.*, 2010) (L_1) and `ncvreg` (Breheny and Huang, 2011) (MCP) to a publicly available throat microbiome dataset (Charlson *et al.*, 2010) including 32 non-smokers and 28 smokers, with the aim to predict the smoking status based on the species relative abundances. Starting from 856 species, we filter out rare species with prevalence less than 10%, and the final list contains 174 species. These 174 species are related to each other based on their genetic similarities as reflected by their patristic distances on the phylogenetic tree constructed from their 16S rRNA gene sequences. A reasonable assumption is that closely related species have similar biological characteristics, hence similar coefficients. We therefore set out to see if we can improve prediction by incorporating the phylogenetic tree information through the graph Laplacian. We construct the Laplacian matrix based on the phylogenetic tree using a similar approach described in Chen *et al.* (2013), where the adjacency coefficients between species are taken to be the inverse of their patristic distances. To compare the prediction performance of the three methods, the samples are randomly divided into two subsets, a training set of half size of samples and a test set of another

Table 1. Prediction accuracy (median/SD) for the throat microbiome dataset

Method	Deviance	AUC	Run time (s)
Glmgraph	1.36(0.24)	0.71(0.08)	4.4
glmnet	1.40(0.11)	0.50(0.10)	0.25
ncvreg	1.40(0.47)	0.63(0.11)	2.4

Note: Computation was conducted under Linux x86_64 system with Intel(R) Xeon(R) CPU X5650 (2.67 GHz).

half of samples. Variable selection and parameter estimation are performed on the training set with 5-fold cross-validation to select the tuning parameters, and the test set is used to evaluate the prediction performance. Fifty random divisions are obtained, and model fitting and testing are carried out on all pairs. We use two criteria, deviance and area under the curve (AUC) to compare all these methods. The result is summarized in Table 1. By incorporating the phylogenetic information, we have achieved better prediction performance for the throat microbiome dataset. We also apply `glmgraph` to a breast cancer gene expression dataset ($n = 286, p = 2563$) to predict the survival outcome based on the gene expression profile. The prediction accuracy is improved by incorporating the gene network information (Supplementary Table S2).

Funding

Gerstner Family Career Development Award.

Conflict of Interest: none declared.

References

Breheny,P. and Huang,J. (2011) Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Ann. Appl. Stat.*, **5**, 232–253.

Charlson,E. *et al*. (2010) Disordered microbial communities in the upper respiratory tract of cigarette smokers. *PLos One*, **5**, e15216.

Chen,J. *et al*. (2013) Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis. *Biostatistics*, **14**, 244–258.

Friedman,J. *et al*. (2010) Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, **33**, 1–22.

Huang,J. *et al*. (2011) The sparse laplacian shrinkage estimator for high-dimensional regression. *Ann. Stat.*, **39**, 2021–2046.

Li,C. and Li,H. (2008) Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, **24**, 1175–1182.

Tian,X. *et al*. (2014) Network-constrained group lasso for high-dimensional multinomial classification with application to cancer subtype prediction. *Cancer Inf.*, **13**, 25–33.

Waldron,L. *et al*. (2011) Optimized application of penalized regression methods to diverse genomic data. *Bioinformatics*, **27**, 3399–3406.