Available online at www.sciencerepository.org

Science Repository

Research Article

Logistic Broken Adaptive Ridge Procedure for Colon Cancer Data Analysis

Hong Yin[1*], Suyun Zhao[2] and Liangzhen Lei[3]

[1]*School of Mathematics, Renmin University of China, Beijing*
[2]*School of Information and Key Lab of Data Engineering and Knowledge Engineering, Renmin University of China, Beijing*
[3]*School of Mathematical Sciences Capital Normal University, China, Beijing*

ARTICLEINFO

Article history:
Received: 16 October, 2019
Accepted: 11 November, 2019
Published: 10 December, 2019

Keywords:
Colon cancer
logistic regression
broken adaptive ridge regression
variable selection
penalty function

ABSTRACT

Background: Colon cancer is the leading cause of cancer-related deaths in the world in both man and women. Knowing the causes and risk factors for colon cancer can help you understanding the importance of routine screening for colon cancer, as well as learn if you are one of the people who should begin screening at the earlier age. Due to the limitation of clinical diagnose, management and treatment outcomes, it is of great necessity to develop effective methods for colon cancer detection and prediction especially cDNA Microarrays and high- density oligonucleotide chips are increasingly used in cancer research.

Methods: Here we propose a novel logistic broken adaptive ridge procedure to address the problem of colon cancer results prediction through selecting effective few variables or genes from 2000 candidate genes.

Results: In total 62 cases with 40 colon cancer patients and 22 healthy patients were included in our analysis. Each case consists of 2000 genes which challenged all the competitive method. From the results, we are so surprised that our proposed method outperforms the classical variable selection approaches in error rate of training data and extra testing data.

Conclusions: Logistic adaptive ridge procedure is very effective for colon cancer predictions, either in terms of prognosis or diagnose. It may benefit patients by guiding therapeutic options. We hope it will contribute to the wider biology and related communities.

Background

Colon cancer (CC) is the third common cause of cancer-related deaths in the world. In 2017, there will be estimated 95520 new cases of colon cancer in the US [1]. While the numbers for colon cancer are fairly equal in men (47700) and women (47820). As of January 1, 2016, there were 724690 men and 727350 women alive in the US with a history of CC [2]. Some of these people were cancer-free, while others still had evidence of cancer and may have been undergoing treatment. Approximately 4.6% of men and 4.2% of women will be diagnosed with CC in their lifetime [1]. The risk of CC increases with age; the median age at diagnosis for colon cancer is 68 in men and 72 in women [3].

In terms of risk factors, a person's change of developing colon cancer increases as he or she gets older, especially after the age of 50. Furthermore, have type 2 diabetes or inflammatory bowel disease, or a family history or colon cancer also increases a person's risk for developing the disease, as do some modifiable risk factors like being overweight and eating a diet rich in red and processed meats [4]. Knowing the causes and risk factors for colon cancer can help you understanding the importance of routine screening for colon cancer, as well as learn if you are one of the people who should begin screening at the earlier age.

In contrast to well-described histopathological findings, data regarding clinical features, management, and treatment outcomes are limited to case reports, hence human being ask for statistical analysis models as auxiliary tools helping doctors to give different prognoses according to various patient's symptoms. For example, patients who died before the considered prognosis period are labeled negative and vice versa.

[*]*Correspondence to: Hong Yin, PhD, Associate Professor of School of Mathematics, Renmin University of China, Beijing, 100872; Tel:(010)82500692; E-mail: yinhong@ruc.edu.cn*

http://dx.doi.org/10.31487/j.COR.2019.5.14

In cancer research, cDNA Microarrays and high-density oligonucleotide chips are increasingly used and, in the meantime, they raise numerous excellent and challenging research problems in fields. By monitoring expression levels in cells for tens of thousands of genes simultaneously, microarray experiments may lead to a better understanding of the molecular variations among tumors and hence to a more informative classification [5]. Over the past years, substantial efforts have been made on gene expression profile using various machine learning techniques for patient outcomes in colon cancer, of which early detection has proven to be important yet challenging. Kop R., et. al [6-9]. Propose a dedicated medical pre-processing pipeline and they claimed that the predictive models generated using their pipeline reconfirmed known predictors and identified new, medically plausible, predictors derived from the cardiovascular and metabolic disease domain, validating the pipeline's effectiveness. Bychkov D. et. al. trained a deep learning-based classifier to predict five-year disease-specific survival in comprehensive series of digitized tumor tissue sample of CRC (colorectal cancer) stained for basic morphology only. Anguraj S analyzed gene expression profiles from 1290 CRC tumors using consensus-based unsupervised clustering and obtained some useful results [7-8]. Some detailed comprehensive molecular characterization of human colon and rectal cancer can be consulted in The Cancer Genome Atlas Network [9]. The Above literatures mentioned are rarely relevant to attributes or variables selection research. Generally, gene expression data is very large attributes but relatively small in samples. Hence, a model including all of the genes is not much parsimonious and explanatory. In this paper, we will focus on variables selection based on penalty function used widely in statistics and our goal is to construct a parsimonious machine learning model and at the same time to let this model inherit high classification accuracy degree.

Table 1: Simulation results of PC and PIC index for variable selection methods.

Method	PC	PIC	Method	PC	PIC
n=200, p=15			n=300, p=50		
Lasso	11.21	0.14	Lasso	44.33	0.29
Adaptive lasso	11.34	0.05	Adaptive lasso	45.59	0.08
Elastic Net	11.49	0.02	Elastic Net	45.24	0.06
SCAD	12.22	0.10	SCAD	46.68	0.11
MCP	12.28	0.09	MCP	47.01	0.08
BAR	12.69	0.09	BAR	47.59	0.11
Logistic BAR	12.78	0.01	Logistic BAR	47.68	0.00

From the above table, we find that the performances of PC and PIC of Logistic BAR technique are the best among all the other variable selection models.

Method

I Binary Logistic Regression

Binary logistic regression is often used for modeling binary outcome variables such as yes "1" or no "0". It is assumed that the binary response, Y, takes the values of 0 and 1 with 0 meaning the trait is not present in observation and 1 meaning the trait is present in observation. Let $X = (X_1, X_2, \dots, X_p)^{\mathrm{T}}$ be a set of explanatory variables. $x_{ij}(i = 1,2,\cdots, j = 1,2,\cdots p)$ is denoted as the observed value of the explanatory variable X_j for the ith observation sample. The binary logistic regression arises from the desire to model the posterior probabilities of the two classes via linear functions in x_{ij}, while at the same time ensuring that they sum to one and remain in $[0,1]$. The model has the form

$$\log\left[\frac{\mathrm{p}}{1-\mathrm{p}}\right] = \beta^{\mathrm{T}}\mathrm{X} = \beta_0 + \beta_1\mathrm{X}_1 + \cdots + \beta_{\mathrm{p}}\mathrm{X}_{\mathrm{p}}$$

where $\mathrm{p} = \Pr(\mathrm{Y} = 1|\mathrm{X} = \mathrm{x})$, $\beta = (\beta_0, \beta_1, \cdots \beta_{\mathrm{p}})^{\mathrm{T}}$. A simple calculation shows that

$$\Pr(\mathrm{Y} = 1|\mathrm{X} = \mathrm{x}) = \frac{\exp(\beta_0 + \beta_1\mathrm{X}_1 + \cdots + \beta_{\mathrm{p}}\mathrm{X}_{\mathrm{p}})}{1 + \exp(\beta_0 + \beta_1\mathrm{X}_1 + \cdots + \beta_{\mathrm{p}}\mathrm{X}_{\mathrm{p}})};$$

$$\Pr(\mathrm{Y} = 0|\mathrm{X} = \mathrm{x}) = \frac{1}{1 + \exp(\beta_0 + \beta_1\mathrm{X}_1 + \cdots + \beta_{\mathrm{p}}\mathrm{X}_{\mathrm{p}})}.$$

Logistic regression models are usually fit by maximum likelihood, using the conditional likelihood of Y given X. So, the likelihood of the binary logistic regression for n observations are

$$\mathrm{L}(\beta) = \prod_{\mathrm{i}=1}^{\mathrm{n}} \Pr(\mathrm{Y}_{\mathrm{i}} = \mathrm{y}_{\mathrm{i}}) = \prod_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{p}^{\mathrm{y}_{\mathrm{i}}}(1-\mathrm{p})^{1-\mathrm{y}_{\mathrm{i}}}$$

And the log-likelihood function is

$$\ln\mathrm{L}(\beta) = \sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{y}_{\mathrm{i}} \log \mathrm{p} + (1-\mathrm{y}_{\mathrm{i}}) \log(1-\mathrm{p}))$$

$$= \sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{y}_{\mathrm{i}} \log \frac{\mathrm{p}}{1-\mathrm{p}} + \log(1-\mathrm{p}))$$

$$= \sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{y}_{\mathrm{i}} \beta^{\mathrm{T}}\mathrm{x} - \log(1 + \exp(\beta^{\mathrm{T}}\mathrm{x}))$$

The above regression model shows that once the regression coefficient β^T is fixed, we can easily compute the probability that $Y = 1$, or the probability that $Y = 0$ for a given observation. To maximize the log-likelihood, we set its derivatives to zero. These score equations are

$$\frac{\partial \ln\mathrm{L}(\beta)}{\partial \beta} = \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{x}_{\mathrm{i}} \left(\mathrm{y}_{\mathrm{i}} - \frac{\exp(\beta^{\mathrm{T}}\mathrm{x})}{1 + \exp(\beta^{\mathrm{T}}\mathrm{x})}\right) = 0$$

which are $\mathrm{p} + 1$ equations nonlinear in β. To solve the score equations above, we use the Newton-Raphson algorithm, which requires the second-derivative or Hessian matrix

$$\frac{\partial^2 \ln\mathrm{L}(\beta)}{\partial \beta\, \partial \beta^{\mathrm{T}}} = -\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{x}_{\mathrm{i}}\mathrm{x}_{\mathrm{i}}^{\mathrm{T}} \mathrm{p}(1-\mathrm{p})$$

Starting with β^{new}, a single Newton-Raphson update is

$$\beta^{new} = \beta^{old} - \left(\frac{\partial^2 lnL(\beta)}{\partial \beta \partial \beta^T}\right)^{-1} \frac{\partial lnL(\beta)}{\partial \beta}.$$

This progress begins with a tentative solution, revises it slightly to see if it can be improved and repeats this revision until no more improvement is made, at which point the process is said to have converged. It is convenient to write the score and Hessian in matrix notation. Let y denote the vector of y_i values, X the $N \times (p+1)$ matrix of x_i values, P the vector of fitted probabilities with ith element $p(x_i; \beta^{old})$ and W an $N \times N$ diagonal matrix of weights with ith diagonal element $p(x_i; \beta^{old})(1 - p(x_i; \beta^{old}))$. Then

$$\frac{\partial \ln L(\beta)}{\partial \beta} = X^T(y - P); \frac{\partial^2 \ln L(\beta)}{\partial \beta \partial \beta^T} = -X^T W X$$

The Newton-Raphson step is thus:

$$\beta^{new} = \beta^{old} + (X^T W X)^{-1} X^T (y - P) = (X^T W X)^{-1} X^T W \left(X\beta^{old} + W^{-1}(y - P)\right)$$

$$= (X^T W X)^{-1} X^T W z$$

where the response $z = X\beta^{old} + W^{-1}(y - P)$ sometimes is known as the adjusted response. These equations get solved repeatedly, since at each iteration P changes, and hence so does W and z. This algorithm is referred to as iteratively reweighted least squares, since each iteration solves the weighted least squares problem:

$$\beta^{new} \leftarrow \text{argmin}_\beta (z - X\beta)^T W (z - X\beta)$$

However, there exists a common and tough question, that is, standard errors of the estimator of coefficient vector β increases when real data sets have a large ratio of variables to cases or high correlations between predictors are very serious. Most researchers think it is the actual final aim to select real and important predictors from candidates rather than only calculate the coefficient estimation values of β. So logistic regression faces a variable selection process.

II Variable Selection Methods

A natural approach to variable selection is L_0-penalized regression, however, the L_0-penalization problem is nonconvex and finding its global optima requires exhaustive combinatorial best subset search, which is NP-hard and computationally infeasible even for data in moderate dimension. Moreover, it can be unstable for variable selection [10]. A popular alternative is L_1-penalized regression, or Lasso, which is known to be consistent for variable selection [11-13]. During the past two decades, much efforts have been devoted to improving Lasso using various variants of the L_1 penalty, which are not only consistent for variable selection, but also consistent for parameter estimation [14-19]. In addition, the L_1-penalized optimization problem can be solved exactly with efficient algorithms and the method became popular since the introduction of the least LASSO method [20]. However, it is known that LASSO does not have the oracle property as it tends to select too many small noise features and is biased for large parameters. In addition, it cannot accommodate the grouping effect when covariates are highly correlated. To address these, several approaches have been proposed including the ALASSO and SCAD [14, 19]. In this article, we propose a broken adaptive ridge (BAR) regression approach that approximates the L0-penalized regression using an iteratively reweighted L2-penalized algorithm for variable selection [21]. The following paragraphs will describe the BAR regression estimation procedure.

III BAR Regression Estimation Procedure

Now suppose that we are interested in simultaneous estimation and variable selection using logistic regression. For this, we will first develop a general penalized estimation procedure, especially, design a new penalty function. To develop a penalized procedure for β based on the estimation procedure mentioned above, it would be very natural to consider and approximate β by

$$g(\tilde{\beta}) = \text{argmin}_\beta \{\frac{1}{2}(z - X\beta)^T W (z - X\beta) + \lambda_n \sum_{j=1}^{p_n} \frac{\beta_j^2}{\tilde{\beta}_j^2}\}$$

Where $\tilde{\beta} = (\tilde{\beta}_1, \tilde{\beta}_2, \cdots, \tilde{\beta}_p)'$ denotes a good initial estimator of β. Then we have

$$g(\tilde{\beta}) = (X^T W X + \lambda_n \sum(\tilde{\beta}))^{-1} X^T W z$$

Where $\sum(\tilde{\beta}) = diag(\tilde{\beta}_1^{-2}, \tilde{\beta}_2^{-2}, \cdots, \tilde{\beta}_p^{-2})$. This suggests we can estimate β by the broken adaptive ridge (BAR) estimator defined as

$$\hat{\beta}^* = \lim_{k \to +\infty} \hat{\beta}^{(k)}$$

Based on the iterated formula $\hat{\beta}^{(k)} = g(\hat{\beta}^{(k-1)})$.

IV Simulation Study

In order to evaluate the quality of various variable selection methods, numerical simulation is needed. We apply the following two criteria to compare the results of each approaches:

1. PC: Percentage of the zeros coefficients that are correctly estimated to be zeros after variable selection;

$$PC = \frac{1}{p - p_0}\left(\frac{1}{n}\sum_{k=1}^{n}\sum_{j=1}^{p} I(\hat{\beta}_{j(k)} = 0) \times I(\beta_j = 0)\right)$$

where p is the number of all the features in the model which will produce the simulation data and p_0 is the selected number of variables from all the features.

2. PIC: Percentage of the non-zeros coefficients that are incorrectly estimated to zeros after variable selection.

$$PIC = \frac{1}{p_0}\left(\frac{1}{n}\sum_{k=1}^{n}\sum_{j=1}^{p} I(\hat{\beta}_{j(k)} = 0) \times I(\beta_j \neq 0)\right)$$

In our simulation, we assume that $X_1, X_2, \ldots, X_n$ and Z are from standard normal distribution $N(0,1)$. Denote covariate $W_i = \frac{(X_i + Z)}{\sqrt{2}} (i = 1,2, \ldots, n)$ and response $Y's$ probability mass function is from

$$Pr(Y = 1|W) = \frac{\exp(W_p + W_p^2 + W_p^3 + 1.5W_q - W_q^2 + 0.7W_q^3)}{1 + \exp(W_p + W_p^2 + W_p^3 + 1.5W_q - W_q^2 + 0.7W_q^3)}$$

where p, q are any two constants less than n. The following table lists two results of scenarios.

An Application

Materials and Results

In this section, we will apply the logistic BAR estimation and variable selection method proposed in the previous sections to two real data sets.

The first data set is about Pima Indian Diabetes. Diabetes is a group of metabolic diseases in which there are high blood sugar levels over a prolonged period. Symptoms of high blood sugar include frequent urination, increased thirst, and increased hunger. To study the reason that leading to diabetes, a cluster of data set about Pima Indian Diabetes was collected. It is consisted of 8 predict variables and 1 response variable (1: Diabetes, 0: Not). The variables are (0):PRG（number of times pregnant）, (1):PLASMA(Plasma glucose concentration in saliva), (2):BP(Diastolic blood pressure), (3):THICK(Triceps skin fold thickness), (4):INSULIN(two Hours serum insulin), (5):BODY(Body mass index: Weight/Height), (6):PEDIGREE(Diabetes pedigree function), (7):AGE (In years). After randomly selecting 768 female patients over 21 years old, 9 variables were taken to fit logistic regression to predict probability that individual females have diabetes. In order to compare the performances of all the variable selection approaches, we randomly select $2/3$ samples from 768 patients as our training data to construct models and remaining $1/3$ samples as our test data to predict their probability that individual females have diabetes or not. Table 2 presents all the results of considered variable selection methods:

Table 2: Variable selection results of Pima Indian Diabetes.

Methods	Variable Number	Variable Name	Accuracy rate of fitting	Accuracy rate of predicting
Logistic	8	[0 1 2 3 4 5 6 7]	0.75390625	0.7421875
Lasso	4	[0 1 5 7]	0.7734375	0.763671875
Adaptive lasso	7	[0 1 2 4 5 6 7]	0.771484375	0.78515625
Elastic Net	5	[0 1 5 6 7]	0.767578125	0.78515625
SCAD	5	[0 1 2 5 6]	0.763671875	0.7890625
MCP	5	[0 1 2 5 6]	0.763671875	0.7890625
BAR	4	[0 1 2 5]	0.79296875	0.79296875
Logistic BAR	5	[0 1 2 5 7]	0.80078125	0.796875

One can see that the results for variable selection number are the same for proposed Lasso and BAR which selected 4 variables and the results of considered Elastic Net, SCAD, MCP and Logistic BAR choose 5 variables which is more one variable than the previous two methods. From accuracy rate of fitting and accuracy rate of predicting, Logistic BAR is the best among other methods. Logistic BAR indicated that skin thickness and two hours serum insulin has not much effect on the occurrence rate of female diabetes. From all the selected variable results, number of times pregnant and Plasma glucose concentration in saliva seemed greatly to related to the occurrence rate of female diabetes.

The second data is about colon cancer which was collected by Alon et. al. and is available at http://microarray.princeton.edu/oncology. This data set consists of 2000 genes measured on 62 patients: 40 diagnosed with colon cancer and 22 healthy patients. The genes are placed in order of descending minimal intensity. Each gene was normalized so its average intensity across the tissues is 0, and its SD is 1. This is a classical data which has a large number of attributes while has a small number of samples. Whatever we want to do with this data set, such as gene clustering, gene classification, the first step requires us to pick up the most important or the most necessary variables from all the candidates variables.

Table 3: The averaged results of variable selection for colon cancer data set over 100 sampling.

Method	Variable number	Error rate of training data	Error rate of testing data
Logistic	2000	1/31	7/31
Lasso	23	2/31	4/31
Adaptive lasso	23	1/31	3/31
Elastic Net	58	1/31	4/31
SCAD	16	1/31	4/31
MCP	11	0/31	3/31
BAR	4	1/31	3/31
Logistic BAR	8	1/31	2/31

In our experiments, we randomly select 20 tumor and 11 normal colon tissues as our training data and the remaining tissues as our testing data. The averaged results over 100 sampling are listed in (Table 3). From the Table 4, Lasso, Adaptive lasso and Elastic Net selected dozens of variables which is far fewer than the original number 2000. Their error rate of testing data is also less than that of logistic regression. SCAD, MCP, BAR and logistic BAR can be partitioned into the same group among which BAR shows the most performance on selecting the number of variables. Although the variables chosen by logistic BAR is 4, the error rate of testing data of logistic BAR is exciting, only 2 misjudgements of 31 samples.

Conclusions

In this paper, we proposed logistic BAR simultaneous variable selection and parameter estimation method for colon cancer and diabetes variable selection and predictions. It is verified that it is a valid and effective method for dealing with high-dimensional gene expression data. Through comparison with classical variable selection method, we show the superiority of logistic BAR. Due to inheriting some good properties of L_0-penalized regression in a sense that it can choose the non-zero components and shrink the zero components quickly, accurately and unhesitatingly. At the same time, it reserves the version of ridge

regression, so there is no much burden in the optimization of objective function. In the feature work, we will make some effort to explore the oracle property and grouping effect of the resulting estimator from the proposed method.

REFERENCES

1. Siegel RL, Miller KD, Jemal (2016) A Cancer statistic 2016. *CA Cancer J Clin* 66: 7-30. [Crossref]
2. Miller KD, Siegel RL, Lin CC, Mariotto AB, Kramer JL et al. Cancer treatment and survivorship statistics 2016. *CA Cancer J Clin* 66: 271-289. [Crossref]
3. Howlader N, Noone AM, Krapcho M (2016) SEER Cancer Statistics Review, 1975-2013. Bethesda, MD: *National Cancer Institute*.
4. Chan DS, Lau R, Aune D, Vieira R, Greenwood DC, Kampman E et al. (2011) Red and processed meat and colorectal cancer incidence: meta-analysis of prospective studies. *PLoS One* 6: e20456. [Crossref]
5. Dudoit S, Fridly J, Speed TP (2002) Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. *J Am Stat Assoc* 97:77-87.
6. Kop R, Hoogendoorn M, Teije AT, Büchner FL, Slottje P et al. (2016) Predictive modeling of colorectal cancer usging a dedicated pre-processing pipeline on routine electronic medical records. *Comput Biol Med* 76: 30-38. [Crossref]
7. Bychkov D, Linder N, Turkki R, Nordling S, Kovanen PE et al. (2018) Deep learningbased tissue analysis predicts outcom in colorectal cancer. *Sci Rep* 8 :3395. [Crossref]
8. Sadanandam A, Lyssiotis CA, Homicsko K, Collisson EA, Gibb WJ et al. (2013) A colorectal cancer classification system that associates cellular phenotype and responses to therapy. *Nat Med* 19: 619-625. [Crossref]
9. Cancer Genome Atlas Network (2012) Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 487: 330-337. [Crossref]
10. Breiman L (1996) Heuristics of instability and stabilization in model selection. *Ann Statist* 24: 2350-2383.
11. Tibshirani RJ (1996) Regression shrinkage and selection via the lasso, JR Stat. Soc. Series B Stat. *Methodol* 58: 267-288.
12. Meinshausen N, Bühlmann P (2006) High-dimensional graphs and variable selection with the lasso. *Ann Statist* 34:1436-1462.
13. Zhao P, Yu B (2006) On model selection consistency of lasso. *J Mach Learn Res* 7: 2541-2563.
14. Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *J Amer Statist Assoc* 96: 1348-1360.
15. Fan J, Xue L, Zou H (2014) Strong oracle optimality of folded concave penalized estimation. *Ann Stat* 42: 819-849. [Crossref]
16. Huang J, Horowitz JL, Ma S (2008) Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Ann Statist* 36: 587-613.
17. Knight K, Fu W (2000) Asymptotics for lasso-type estimators. *Ann Statist* 28: 1356-1378.
18. Zhang C, H Nearly (2010) unbiased variable selection under minimax concave penalty. *Ann Statist* 38: 894-942.
19. Zou H (2006) The adaptive lasso and its oracle properties. *J Amer Statist Assoc* 101: 1418-1429.
20. Tibshirani R (1996) Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B 58: 267-288.
21. Dai L, Chen K, Sun Z, Liu Z, Li G (2018) Broken adaptive ridge regression and its asymptotic properties. *J Multivar Anal* 168: 334-351. [Crossref]