Bio-Medical Materials and Engineering 24 (2014) 3447–3454
DOI 10.3233/BME-141169
IOS Press

Application of $L_{1/2}$ regularization logistic method in heart disease diagnosis

Bowen Zhang, Hua Chai, Ziyi Yang, Yong Liang*, Gejin Chu and Xiaoying Liu
Faculty of Information Technology & State Key Laboratory of Quality Research in Chinese Medicines, Macau University of Science and Technology, Taipa 999078, Macau, China

Abstract. Heart disease has become the number one killer of human health, and its diagnosis depends on many features, such as age, blood pressure, heart rate and other dozens of physiological indicators. Although there are so many risk factors, doctors usually diagnose the disease depending on their intuition and experience, which requires a lot of knowledge and experience for correct determination. To find the hidden medical information in the existing clinical data is a noticeable and powerful approach in the study of heart disease diagnosis. In this paper, sparse logistic regression method is introduced to detect the key risk factors using $L_{1/2}$ regularization on the real heart disease data. Experimental results show that the sparse logistic $L_{1/2}$ regularization method achieves fewer but informative key features than Lasso, SCAD, MCP and Elastic net regularization approaches. Simultaneously, the proposed method can cut down the computational complexity, save cost and time to undergo medical tests and checkups, reduce the number of attributes needed to be taken from patients.

Keywords: Heart disease, feature selection, sparse logistic regression, L1/2 regularization

1. Introduction

The diagnosis of heart disease in most cases depends on a complex combination of clinical and pathological data. This complexity leads to excessive medical costs affecting the quality of the medical care [1]. According to the world health statistics 2012 report, compiled by World Health Organization (WHO), cardiovascular diseases, such as heart failure, heart attack or stroke, cause 17.5 million deaths every year, nearly one third of the total deaths in the world [2]. In the last two decades, many risk factors have been collected for the heart disease, such as age, sex, blood pressure, and heart rate. However, it is complicated to analyze heart disease with patients' clinical data. Particularly, doctors usually diagnose illness depending on their intuition and experience, rather than on the significant data hidden in the database. It requires a lot of knowledge and experiences for correct diagnosis. Therefore, finding the key risk factors will be of tremendous assistance to diagnosis.

The detection of biomarkers in heart disease is one of the key contributions using computational biology. Due to the wide availability of huge amounts of data, data mining has been an important method in the information industry and there is a need for turning such data into useful information. Data mining techniques have been applied to a variety of medical domains to improve medical diagnosis [3].

*Corresponding author: Yong Liang, Faculty of Information Technology, Macau University of Science and Technology, Macau, China. Tel.: 0853-88972034; Fax: 0853-2882 3280; Email: yliang@must.edu.mo.

To find the hidden medical information from different expressions between the healthy individuals and individuals with heart disease in the existing clinical data is a noticeable and powerful approach in the study of heart disease classification [4]. Logistic regression is widely used in machine learning for classification problems. In the last two decades, the regularization approaches have been developed to avoid over-fitting for logistic regression, especially when there is only a small number of training examples, or when there are a large number of parameters to be learned. In particular, regularized logistic regression is often used for feature selection, and was shown to have good generalization performance in the presence of many irrelevant features.

In this paper, sparse logistic regression method is introduced to detect the key risk factors using $L_{1/2}$ regularization on the heart disease datasets. In fact, nonconvex regularization methods associated with the Lq ($0<q<1$) norm can find very sparse solutions. Moreover, the $L_{1/2}$ penalty can be somehow regarded as a representation among all the Lq ($0<q<1$) penalization and enjoy many attractive statistical properties, such as the globally necessary optimality condition [5]. Therefore, to solve the $L_{1/2}$ regularized logistic approach concerning the heart disease diagnosis, a coordinate descent algorithm is applied. Experimental results show that the sparse $L_{1/2}$ logistic method attains similar prediction accuracy compared with Lasso [6], SCAD [7], MCP [8] and Elastic net [9] regularization approaches. Moreover, a greater advantage of the $L_{1/2}$ logistic method is that fewer but informative key features are selected for heart disease diagnosis problem.

2. Sparse $L_{1/2}$ regularization logistic method

Suppose there are n samples, $D=\{(X_1, y_1),(X_2, y_2),....,(X_n, y_n)\}$, where $X_i=(x_{i1}, x_{i2},..., x_{ip})$ is i^{th} input instance and consists of p features, x_{ij} is the value of feature j and y_i is the class indicator for the i^{th} instance. A classier $f(x)=e^x/(1+e^x)$ is defined so that for any input x with class indicator y, $f(x)$ can predict y accurately. The logistic regression is written as follows:

$$P(Y_i=1 \mid X_i)=f(X_i^{'}\beta)=\frac{\exp(X_i^{'}\beta)}{1+\exp(X_i^{'}\beta)} \tag{1}$$

where $\beta=(\beta_0, \beta_1,..., \beta_p)$ are the corresponding coefficients and they are able to be estimated. Note that β_0 is the intercept. The log likelihood function of logistic regression is expressed as:

$$l(\beta \mid D)=-\sum_{i=1}^{n}\{y_i \log[f(X_i^{'}\beta)]+(1-y_i)\log[1-f(X_i^{'}\beta)]\} \tag{2}$$

In practice, the evaluation measurement for the quality of a model will differ due to the circumstances. Typically two situations are considered: (*i*) prediction accuracy on future data. It is difficult to defend a model that predicts poorly; (*ii*) model's interpretation. Researchers prefer a simple model

because it throws light on the relationship between the response and covariates more clearly. When the number of predictors is large, parsimony is especially an important issue. It is well known that the logistic regression often does poorly in both interpretation and prediction. Therefore, the penalization techniques for logistic regression have been proposed as:

$$\beta = \arg\min\{l(\beta \mid D) + \lambda \sum_{j=1}^{p} P(\beta_j)\} \tag{3}$$

where $P(\beta)$ is a penalized function and $\lambda > 0$ is used to balance the effects of the two parts, which could be chosen according to properties of the two norms, or can be tuned empirically. Lasso (L_1 penalty) is a very popular regularization technique in the recent years, which has penalty function $P(\beta) = \sum|\beta|$. Moreover, many L_1 type regularizations have also been proposed for logistic regression, such as Lasso, SCAD, MCP and Elastic net regularization approaches. However, L_1 type penalization may not yield sufficiently sparse variable selection in real applications.

In fact, Lq ($0<q<1$) norm regularization $P(\beta) = \sum|\beta|^q$ of q, which is a low value, would obtain more spare solutions. However, q is very close to zero, which leads to some difficulties with the increase of convergence. So the characteristics of Lq regularization were explored. This paper applied the $L_{1/2}$ regularization of extreme importance. When $1/2 \le q \le 1$, obviously, the result of the $L_{1/2}$ regularization is most sparse. Moreover, compared with the L_1 regularization, the convergence complexity of $L_{1/2}$ regularization algorithm is not very high. Meanwhile the ability of feature selection shows little difference between the $L_{1/2}$ regularization and the Lq regularization as $0 < q < 1/2$. Consequently, the $L_{1/2}$ regularization is an important and special representative in Lq ($0 \le q \le 1$) regularizations [5]. In this paper, the $L_{1/2}$ penalized function was employed to the log-likelihood function of the logistic regression model. Therefore, the $L_{1/2}$ regularized logistic regression was expressed as:

$$\beta = \arg\min\{l(\beta \mid D) + \lambda \sum_{j=1}^{p} |\beta_j|^{1/2}\} \tag{4}$$

There are many attractive properties in the $L_{1/2}$ regularization, such as the globally necessary optimality condition, oracle, unbiasedness, sparsity properties and so on. The $L_{1/2}$ regularization is an innovative method by lots of theoretical and experimental analyses. Our work in this paper also demonstrated the effectiveness of the $L_{1/2}$ regularized logistic regression with a small number of relevant features for heart disease diagnosis.

3. A coordinate descent algorithm for the L1/2 regularization logistic model

The penalty function of $L_{1/2}$ regularization is nonconvex, which raises numerical challenges in fitting the logistic model. Recently, coordinate descent algorithms [10] for solving nonconvex models (SCAD, MCP) have been shown significant efficiency and convergence [11]. The algorithms optimize a target function with respect to a single parameter at a time, iteratively cycling through all parameters

until reaching its convergence. Since the computational burden increases only linearly with p, coordinate descent algorithms can be a powerful tool for solving high-dimensional problems.

Therefore, in this paper, a novel univariate half thresholding operator of the coordinate descent algorithm is proposed for the $L_{1/2}$ regularization, which can be written as follows:

$$\beta_j = New\_Half(\omega_j, \lambda) = \begin{cases} \frac{2}{3}\omega_j\left(1+\cos\left(\frac{2(\pi-\varphi_\lambda(\omega_j))}{3}\right)\right) & if\ |\omega_j| > \frac{\sqrt[3]{54}}{4}(\lambda)^{\frac{2}{3}} \\ 0 & otherwise \end{cases} \tag{5}$$

where $\tilde{y}_i^{(j)} = \sum_{k\neq j} x_{ik}\beta_k$ as the partial residual for fitting β_j, and $\omega_j = \sum_{i=1}^{n} x_{ij}\left(y_i - \tilde{y}_i^{(j)}\right)$, where $\varphi_\lambda(\omega) = \arccos\left(\frac{\lambda}{8}\left(\frac{|\omega|}{3}\right)^{-\frac{3}{2}}\right)$

.

Remark: In our previous work [12], $\frac{3}{4}(\lambda)^{\frac{2}{3}}$ was used for representing $L_{1/2}$ regularization thresholding operator. Here, a new half thresholding representation $\frac{\sqrt[3]{54}}{4}(\lambda)^{\frac{2}{3}}$ was proposed. This new value is more precise and effective than the previous one. Since it is known that the quantity of the solutions of a regularization problem seriously depends on the setting of the regularization parameter λ. Based on this novel thresholding operator, when λ is chosen by some efficient parameters tuning strategy, such as cross-validation, the convergence of the algorithm is proved [13].

Algorithm: The coordinate descent algorithm for the $L_{1/2}$ regularization logistic method
Step 1: Initialize all $\beta_j(m) = 0$ (j =1,2,...,p) and set m = 0, λ chosen by cross − validation;
Step 2: Approximate the loss function (3) based on the current $\beta(m)$;
Step 3: Update each $\beta_j(m)$, and cycle over j=1,…,p, until $\beta_j(m)$ does not change;
Step 3.1: Calculate $\tilde{y}_i^{(j)}(m) = \sum_{k\neq j} x_{ik}\beta_k$ and $\omega_j(m) = \sum_i^n x_{ij}\left(y_i(m) - \tilde{y}_i^{(j)}(m)\right)$;
Step 3.2: Update $\beta_j(m) = New\_Half(\omega_j(m), \lambda)$;
Step 4: Let $m = m + 1$, $\beta(m+1) \leftarrow \beta(m)$; Repeat Steps 2, 3 until the convergence of $\beta(m)$.

The coordinate descent algorithm was presented to the $L_{1/2}$ regularization logistic approach resulting in improvements of convergence speed in practice. The results are consistently robust throughout the convergence process. This improved performance may help reduce computational time to reach a desired result of feature selection, or improve the quality of classification and prediction for heart disease diagnosis with limited computational time.

Table 1

Three cardiology patient datasets used in the experiments

Name of Datasets	Original Datasets		Datasets after preprocessed	
	Total instances	Attributes	Total instances	Attributes
Cleveland	303	76	270	46
Hungarian	294	76	245	37
Long Beach VA	200	76	103	50

4. Simulation and discussion

4.1. Description of datasets

These cardiology patient datasets are collected at the University of California, Irvine (UCI). The aim of these datasets is to classify the presence or absence of heart disease given the results of various medical tests carried out on a patient. These datasets together with one or more data mining techniques can be used to help us develop profiles for differentiating individuals with heart disease from those with unknown heart conditions. Table 1 presents the statistical comparison of the three cardiology patients' datasets. For example, the initial Cleveland data set had a total of 303 instances and 76 clinical features. After data preprocessing, 151 out of 270 instances belonged to healthy and 119 instances belonged to the heart disease, in the meantime, 46 clinical features had been recorded for each instance.

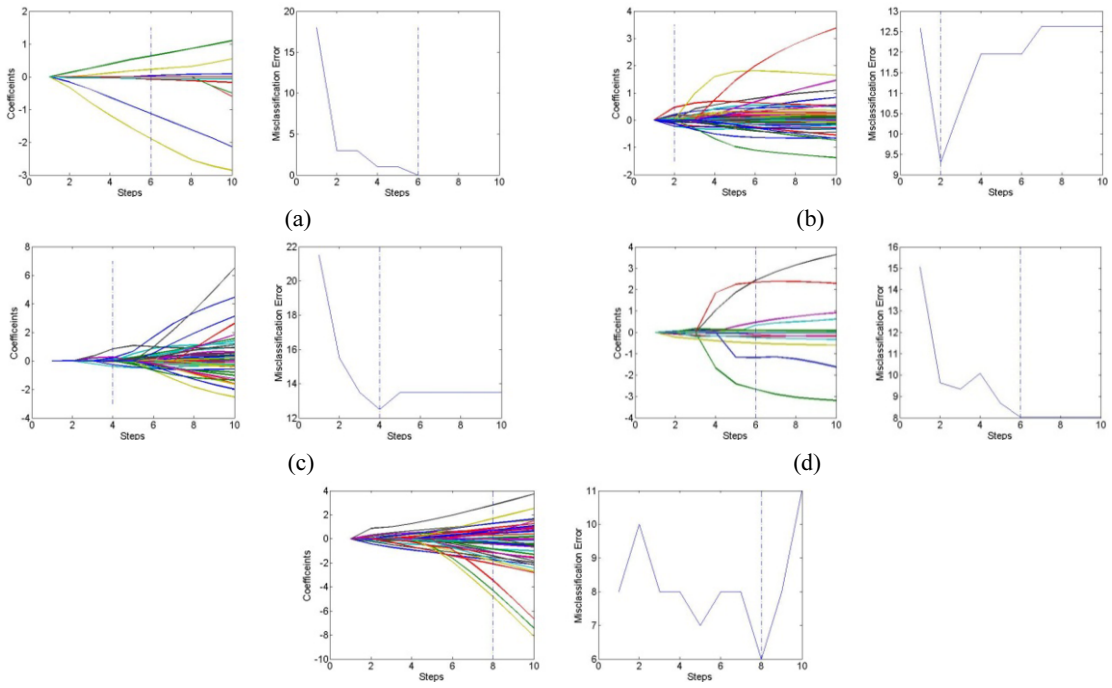

Fig. 1. The regularization paths and the misclassification errors of the five approaches for the Cleveland data set in the singl run. (a) $L_{1/2}$ regularization, (b) Lasso, (c) SCAD, (d) MCP and (e) Elastic net.

Table 2

The averaged key feature selected and the averaged prediction accuracies of five regularization methods for three cardiology patient datasets

Methods	Averaged key features selected			Averaged prediction accuracies		
	Cleveland	Hungarian	Long Beach VA	Cleveland	Hungarian	Long Beach VA
$L_{1/2}$	5	11	9	76.8%	86.8%	81.6%
Lasso	16	16	16	78.1%	85.9%	80.4%
SCAD	14	11	11	75.6%	86.1%	81.2%
MCP	11	13	13	76.3%	85.6%	81.8%
Elastic net	31	23	25	79.4%	86.4%	80.1%

4.2. Experiments and results

In this section, the simulation performance of different methods is compared with regard to a) prediction error, b) the number of non-zero coefficients in the model, and c) "misclassification error" of the variables retained. The methods compared areL1/2 regularization, Lasso [6], SCAD [7], MCP [8] and Elastic net [9,14].

Figures 1(a)-1(e) display the coefficient paths and the misclassification errors corresponding to selected features by the five regularization methods respectively. The vertical dotted line is drawn at the optimal solution, which can be determined by the minimal misclassification errors based on the ten-fold cross validation. The features selected by the $L_{1/2}$ regularization, Lasso, SCAD, MCP and Elastic net methods are 5, 31, 16, 14 and 11 respectively. The results of the feature selection show that our proposed $L_{1/2}$ method is superior to other four regularization methods.

To further compare the predictive accuracies of the five regularized logistic approaches, caused by the heart disease datasets, approximately 60% of the samples are used for the training set and the other 40% for the test set, which are drawn randomly without replacement. The 10-fold cross validation (CV) approach is used on the training sets to tune the regularization parameter λ. Table 2 presents the averaged key feature selected and the averaged prediction accuracies, which are averaged over 50 runs. In the Cleveland dataset, the classification model with about 5 features selected by the $L_{1/2}$ regularization achieves the averaged prediction accuracy 76.8%, while the classifiers with Lasso, SCAD, MCP and Elastic net methods show averaged prediction accuracies of 78.1%, 75.6%, 76.3% and 79.4% with about 16, 14, 11, and 31 features selected respectively. These obtained results prove that the selected discriminative features for classification have indeed improved the performance of the classifier as the regularization methods achieve similar accuracies for the Cleveland datasets. Note that, just roughly 5 features are used by the $L_{1/2}$ method, which are much smaller than 16, 14, 11, 31 obtained by other regularization methods. For Hungarian dataset, the $L_{1/2}$ method outperforms other four regularization methods in terms of prediction accuracy and variable selection. For Long Beach VA dataset, the $L_{1/2}$ regularization method achieves better classification performance than the Lasso, SCAD, and Elastic net methods and is only worse than the MCP method. However, the feature selection properties of the $L_{1/2}$ method are better. These results have shown that the $L_{1/2}$ regularized regression method is effective and robust in the classification of heart disease datasets, because the development of the simple and low-cost test is important for screening and diagnostic applications.

Table 3 shows the features commonly selected by all five regularization methods. They may indicate the most relevant features for heart disease diagnosis. For example, sex and exercise-induced angina are selected by all five methods in two different datasets, which means that enough attention should be paid to the two features. Besides, some features are only discovered by the L1/2 approach,

Table 3

The common key features selected by all five regularization methods for three cardiology patients' datasets

Datasets	Common key features selected
Cleveland	Sex, Day of exercise ECG reading, Nitrates used during exercise ECG, Exercise-induced angina
Hungarian	Sex, Provoked by exertion, Relieved after rest, Resting blood pressure, Exercise-induced angina, ST depression induced by exercise relative to res
Long Beach VA	Provoked by exertion, Number of years as a smoker, Fasting blood sugar > 120 mg/dl, Day of cardiac cath, Cxmain and rcaprox

but are not selected by other regularization methods. For example, the feature height at rest in Hungarian dataset and the feature ramus in Long Beach VA dataset are only selected by the L1/2 regularization method.

5. Conclusion

In this paper, a heart disease prediction system is proposed to filter features without significant contributions to a given high-level diagnosis. The sparse logistic regression method is introduced to detect the key risk factors using $L_{1/2}$ regularization on the heart disease datasets. Then the proposed $L_{1/2}$ regularization logistic method is compared with other regularization approaches. Experimental results show that the $L_{1/2}$ regularization achieves fewer but informative key features than Lasso, SCAD, MCP and Elastic net regularization approaches. This research reveals that regularization logistic approach helps increase computational efficiency while improving classification and prediction accuracy. Simultaneously, it can save cost and time to undergo medical tests and checkups, ensuring that the patient can monitor his health on his own and plan preventive measures and treatment at early stages. According to our experimental results, the attributes *sex* and *exercise-induced angina* are demonstrated to be the most relevant features of heart disease.

Acknowledgment

The work described in this paper was partially supported by STDF of Macau 099/2013/A3 and the NSFC projects (Grant No. 11131006 and No. 11171272).

References

[1] B. Fida, M. Nazir, N. Naveed and S. Akram, Heart disease classification ensemble optimization using genetic algorithm, Proceeding of IEEE 14th International Multitopic Conference, 2011, 19–24.
[2] A. Khemphila and V. Boonjing, Heart disease classification using neural network and feature selection, Proceeding of 21st International Conference on Systems Engineering, 2011, 406–409.
[3] E.L. Leung, Z.W. Cao, Z.H. Jiang, H. Zhao and L. Liu. Network-based drug discovery by integrating systems biology and computational technologies, Briefings in Bioinformatics **14** (2013), 491–505.
[4] A.H. Chen, S.Y. Huang, P.S. Hong, C.H. Cheng and E.J. Lin, HDPS: heart disease prediction system, Computing in Cardiology **38** (2011), 557–560.
[5] Z.B. Xu, H. Zhang, Y. Wang, X.Y. Chang and Y. Liang, $L_{1/2}$ regularization, Sci. China Series F **40** (2010), 1–11.
[6] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Statist. Soc. B **58** (1996), 267–288.

[7] J. Fan and R. Li, Variable selection for Cox's proportional hazards model and frailty model, Ann. Statist. **30** (2002), 74–99.
[8] C.H. Zhang, Nearly unbiased variable selection under minimax concave penalty, Institute of Mathematical Statistics **38** (2010), 894–942.
[9] H. Zou and T. Hastie, Regularization and variable selection via the elastic net, Proceedings of Royal Statistical Society **2** (2005), 301–320.
[10] J. Friedman, Regularization paths for generalized linear models via coordinate descent, J. Stat. Softw. **33** (2010), 1–22.
[11] P. Breheny and J. Huang, Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection, Ann. Appl. Stat. **5** (2011), 232–253.
[12] Y. Liang et al., Sparse logistic regression with a $L_{1/2}$ penalty for gene selection in cancer classification, BMC Bioinformatics **14** (2013), 198.
[13] Z.B. Xu et al., $L_{1/2}$ regularization: a thresholding representation theory and a fast solver, IEEE Transactions on Neural Networks and Learning Systems **23** (2012), 1013–1027.
[14] A. Liu, Z. Gao, T. Hao, Y.T. Su and Z.X. Yang, Sparse coding induced transfer learning for HEp-2 cell classification, Bio-Medical Materials and Engineering **24** (2014), 237–243.