Split and Conquer Method in Penalized Logistic Regression with Lasso (Application on Credit Scoring Data)

F Shofiyah\(^1\), and A Sofro\(^{1,*}\)

\(^1\)Mathematics Department, Universitas Negeri Surabaya, Ketintang, Surabaya 60231, East Java, Indonesia

\(^*\)Corresponding Author’s Email: ayuminsofro@unesa.ac.id

Abstract. Big data is one of the biggest issues recently. We need a new approach to deal with the problem. One of a statistical strategy that can be used to solve the problem is split and conquer method. In this paper, we focus on non-Gaussian data, i.e binomial distribution. In this research will be discussed about the implementation of the method for credit scoring data. The result is there are 5 important independent variables in credit scoring data. The first variable is the percentage of total balances in credit cards and private lines of credit except real estate divided by the number of credit limits. The second variable is the age of debtor. The third variable is how many times debtor has been 30-59 days late pay in the last 2 years. The fourth variable is how many times the debtor has been late for pay 90 days or more. The last variable is how many times debtor has been 60-89 days late pay in the last 2 years.

1. Introduction
Big data is datasets that so voluminous and complex, so it is difficult to be analyzed. A statistical strategy that can be used to solve big data problems is split and conquer method [1]. The method solves the big data problem by divide big data into Q blocks, so it can be easily processed by computing. The estimator results of each block are combined in the final estimate. One example of data that requires big data analysis is credit scoring data [2].

Credit scoring is the assessment process using the historical data of the debtor [2]. The historical data is used to classify debtors who are eligible to be loaned and not. Credit scoring data can contain millions or even billions of debtor’s historical records. If all data is used in the classification model, then it will be very inefficient and time-consuming [2]. For that reason, we can use the Split and Conquer method. In addition, credit scoring data requires selecting variables to identify relevant variables. The method that can be used is penalized logistic regression. Penalized logistic regression is a logistic regression that estimates coefficients by maximizing log-likelihood function plus the penalty function on coefficient size [3].

One interesting penalty function is the Least Absolute Shrinkage Selection Operator (LASSO). LASSO was proposed by Tibshirani in 1996, see [4]. Efron (2004) has been introduced a regression scheme that involved LASSO as one in it, sees [5]. LASSO penalty function will able to generate regression with a good parameter estimator and further minimize the error from the logistic regression model [6]. LASSO penalty function also plays a role in identifying which predictor variables are important for the response variable. LASSO penalty reduces the number of coefficients by making it 0. Thus LASSO can reduce the dimensions of the data credit scoring.
In this research will be analyzed big data on real secondary data credit scoring use the Split and Conquer method on penalized logistic regression with LASSO penalty. The credit scoring data consisted of 150,000 observations, 1 dependent variable dependent, and 10 independent variables.

2. Method
2.1 Logistic Regression
The logistic regression model is a model that describes the relationship between several factors (predictor variables) with dichotomous (binary) response variables. For the response variable, \( y = 1 \) to express the results of a successful and \( y = 0 \) express the result of the failed. In logistic regression, the variable response is Bernoulli distributed. The response variable is usually denoted by \( y \) and the predictor variable is symbolized by \( x \). Suppose \( y_i \) is the \( i \)-th response variable, \( \pi_i \) is the \( i \)-th probability of success, and \( i \) is the observation measure, the probability function of \( y \) is:

\[
f(y_i, \pi_i) = \begin{cases} 
\pi_i (1 - \pi_i)^{1-y_i}, & \text{for } y_i = 0, 1 \\
0, & \text{otherwise} 
\end{cases}
\]  

The logistic model with \( p \) predictor variable is expressed as follows

\[
\pi_i = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)}. 
\]  

The logistic regression model in equation (2) is transformed using the logit transformation obtained:

\[
\log \left( \frac{\pi_i}{1-\pi_i} \right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p. 
\]  

To estimate logistic regression, parameter estimation method is required, that is maximum likelihood estimation

2.2 Parameter Estimation
The Maximum Likelihood Estimation / MLE method is one of the methods in parameter estimation. This method was first introduced by R.A Fisher in 1912, see [7]. Let \( y_1, y_2, \ldots, y_n \) be a continuous random variable of size \( n \) with probability functions \( f(y_i; \beta) \) and \( \beta \) are unknown parameters. In logistic regression, the likelihood function is as follows [8]

\[
L(y; \beta) = \prod_{i=1}^{n} [\pi_i^{y_i}(1 - \pi_i)^{1-y_i}], 
\]  

and the likelihood function is as follows [8]

\[
\log L(y; \beta) = \sum_{i=1}^{n} \{y_i \log(\pi_i) + (1 - y_i) \log(1 - \pi_i)\}. 
\]  

Parameter estimation (\( \beta \)) can be obtained by maximizing log-likelihood function.

2.3 Least Absolute Shrinkage and Selection Operator (LASSO)
LASSO method was first introduced by Tibshirani in 1996, see [4]. LASSO shrinks regression coefficients from predictor variables with high correlation with errors, being either zero or near zero [8]. The LASSO penalty function is

\[
P(\beta) = \sum_{j=1}^{p} |\beta_j|, 
\]  

with \( \sum_{j=1}^{p} |\beta_j| \leq \lambda \), and \( \lambda \) is a tuning parameter that controls the depreciation of the LASSO coefficient by \( \lambda \geq 0 \). LASSO can shrink some coefficients toward zero even exactly zero [6], so it can select predictor variables.

2.4 Penalized Logistic Regression Model with LASSO penalty
There are several practical problems and computational problems in the regression analysis, such as multicollinearity and high dimension. To deal with this problem, we need a variable selection method. Penalized regression is a modern approach to choosing a regression model with high prediction accuracy and efficiency [7]. The penalized regression has the advantage of produce more stable results in highly correlated data and data that has a much larger number of predictors than the sample size. The log-likelihood of Penalized logistic regression use LASSO penalty on equation (6) is as follows [9]

\[
\log L_p (\mathbf{y}; \mathbf{\beta}) = \sum_{i=1}^{n} -[y_i \log(\pi_i) + (1 - y_i) \log(1 - \pi_i)] + \lambda \sum_{j=1}^{p} |\beta_j|. \quad (7)
\]

Some coefficients will be equal to zero, depend on LASSO penalty. Therefore, the LASSO performs the selection of variables.

The LASSO coefficient estimate is as follows [9]

\[
\hat{\mathbf{\beta}}_{\text{LASSO}} = \arg \min_{\mathbf{\beta}} \left\{ -\sum_{i=1}^{n} \{y_i \log(\pi_i) + (1 - y_i) \log(1 - \pi_i)\} + \lambda \sum_{j=1}^{p} |\beta_j| \right\}. \quad (8)
\]

After estimating, then do the optimization through coordinate descent.

2.5 Optimization on Penalized Logistic Regression using Coordinate Descent

Suppose the current estimates of the parameter are \((\hat{\beta}_0, \hat{\mathbf{\beta}})\). To make calculations easier, we form the approximation of the squared log-likelihood regression penalized logistics use the Taylor expansion [10], i.e.

\[
(\hat{\beta}_0, \hat{\mathbf{\beta}}) = \arg \min_{\mathbf{\beta}} \frac{1}{n} \sum_{i=1}^{n} w_i (z_i - \beta_0 - x_i^T \hat{\mathbf{\beta}})^2 + \lambda \sum_{j=1}^{p} |\beta_j|, \quad (9)
\]

where

\[
\hat{\eta}_i = \hat{\beta}_0 + x_i^T \hat{\mathbf{\beta}}; \quad \tilde{\eta}_i = \frac{1}{1 + \exp(-\eta_i)}; \quad z_i(\tilde{\eta}_i) = \tilde{\eta}_i + \frac{y_i - \tilde{\eta}_i}{\tilde{\eta}_i(1 - \tilde{\eta}_i)}; \quad w_i(\tilde{\eta}_i) = \tilde{\eta}_i(1 - \tilde{\eta}_i).
\]

The solution for equation (9) is

\[
\hat{\beta}_j = \frac{S \left( \frac{1}{n} \sum_{i=1}^{n} w_i x_{ij} \tilde{\eta}_i + \frac{1}{n} \sum_{i=1}^{n} w_i x_{ij}^2 \tilde{\eta}_i \hat{\beta}_j + \lambda \alpha \right)}{\frac{1}{\lambda^{\alpha}} \sum_{i=1}^{n} w_i x_{ij}^2 + \lambda (1 - \alpha)}. \quad (10)
\]

In the equation (10) there is a function \(S\). The function \(S\) is soft-thresholding which has the following value

\[
S(c, y) = \begin{cases} 
  c - y, & \text{if } c > 0, \text{and } y < |c| \\
  c + y, & \text{if } c < 0, \text{and } y < |c| \\
  0, & \text{if } y \geq |c|
\end{cases} \quad (11)
\]

The following will explain the penalized likelihood optimization algorithm on penalized logistic regression use coordinate descent [11]

1. Initialization \((\hat{\beta}_0, \hat{\mathbf{\beta}})\), set \(\tilde{\eta}_i = \hat{\beta}_0 + x_i^T \hat{\mathbf{\beta}}\) for \(i = 1, 2, ..., n\)
2. Calculate \(z_i = z_i(\tilde{\eta}_i)\) and \(w_i = w_i(\tilde{\eta}_i)\)
3. Estimate \(\beta\) with equation (9)
4. for \(j = 1, 2, ..., p\) and \(r_i = y_i - \hat{\beta}_0 - x_i^T \hat{\mathbf{\beta}}\), calculate the solution \(\hat{\beta}_j\) with the equation (10)
5. Set \(\hat{\beta} = \hat{\beta}, \hat{\beta}_0 = \hat{\beta}_0\)
6. Repeat steps 2-4 until \(\beta\) convergence
The coordinate descent algorithm requires tuning parameters, then we will look for initial value tuning parameters use pathwise solution

2.6 Pathwise Solution
To select the value of \( \lambda \), we first choose the grid/value range of \( \lambda \), that is \( \lambda_{\text{max}} \) and \( \lambda_{\text{min}} \) value. The first step is to find \( \lambda \) by setting \( \beta = 0 \) [12]. To find the value of \( \lambda_{\text{max}} \) is as follows:

\[
\lambda_{\text{max}} = \max_j \frac{1}{n \epsilon} \sum_{i=1}^n w_i(0) x_{ij} z(0)_i, \tag{12}
\]

where \( \lambda_{\text{min}} = \epsilon \lambda_{\text{max}} \). The fixed value for \( \epsilon \) depends, for \( n < p \), then \( \epsilon = 0.05 \) and for \( n \geq p \), then \( \epsilon = 0.0001 \). After it will be selected optimal tuning parameters using k-cross-validation

2.7 k-fold Cross Validation (KCV)
In k-fold cross-validation, the datasets are divided randomly into the same size set, then find for cross-validation value \( \lambda \), i.e.

\[
CV(\lambda) = \ell(\beta_{-\lambda}(\lambda)) - \ell_{-i}(\beta_{-\lambda}(\lambda)). \tag{13}
\]

Where \( \ell_{-i} \) is the partial log of likelihood data block \( k-1 \) (except the i-th data block).

Our total goodness of fit estimate, \( CV(\lambda) \), is the sum of all \( CV_i(\lambda) \). Then, choose \( \lambda \) values that maximize cross-validation \( CV(\lambda) \).

2.8 Split and Conquer Methods for Penalized Regression with LASSO on Big Data
Split and Conquer method developed by Chen and Xie in 2014, see [13]. This method has advantages that are to overcome the problem of data that is too large.

\( n \) data is divided into \( Q \) data block, then we find the parameter estimation from each data block uses the penalized logistic regression with LASSO.

In this section will be discussed about the conquer method, which combines the value of \( \hat{\beta} \) of each data block. The combined estimate of with \( q = 1, \ldots, Q \) is

\[
\hat{\beta}^{(c)} = A \left( \sum_{q=1}^Q A^T S_q A \right)^{-1} \sum_{q=1}^Q A^T S_q A \hat{\beta}_{q, \mathcal{A}_q}. \tag{14}
\]

The following will explain how to calculate each of the above variables:

Defined

\[
\mathcal{A}_q = \{ j : \hat{\beta}_{q,j} \neq 0 \}. \tag{15}
\]

With \( j \) is index of variable, so \( j = 1, 2, \ldots, p \)

\( \hat{\beta}_{q,j} \) is the estimator of the \( q \)-th data block on the \( j \)-th variable. \( \mathcal{A}_q \) is the set of an index the non-zero estimator. Thus, \( \hat{\beta}_{q, \mathcal{A}_q} \) is a sub-vector contain only the nonzero estimator [13].

Defined

\[
\mathcal{A}^{(c)} = \left\{ j : \sum_{q=1}^Q 1(\hat{\beta}_{q,j} \neq 0) > \omega \right\}. \tag{16}
\]

\( \mathcal{A}^{(c)} \) is the set of selected variables index of the combined estimator, where \( \omega \in [0, Q) \) is a predefined threshold and 1 is an indicator function [13]. So if the index \( j \) appears more than \( \omega \) times in \( \mathcal{A}_q \), then the \( j \) index is included to the set \( \mathcal{A}^{(c)} \).

So, \( \hat{\beta}_{q, \mathcal{A}^{(c)}} \) is the estimator of the variable on \( q \)-th data block and \( j \)-th variables, where \( j \) is \( j \) which belongs to the set \( \mathcal{A}^{(c)} \).

Suppose \( E = \text{diag}(v_1, \ldots, v_p) \) is a \( p \times p \) matrix with \( v_j = 1 \) if \( \sum_{q=1}^Q 1(\hat{\beta}_{q,j} \neq 0) > \omega \) and 0 if otherwise, and \( A = E_{\mathcal{A}^{(c)}} \) is the matrix \( p \times |\mathcal{A}^{(c)}| \), ie the submatrix of the matrix \( E \), where the rows
chosen from $E$ is the rows $j$ contained in the set $\mathcal{A}^{(c)}$. Defined $S_q = X_q^T \sum(\hat{\theta}_q)X_q$ where $\hat{\theta}_q = X_q \hat{\beta}_q$ and $\sum(\hat{\theta}_q) = \text{diag}(\sigma(\theta_{i1}), ..., \sigma(\theta_{iQ}))$. The value of $\sigma_i = b"(\theta_i) = \frac{e^{\theta_i}}{(1+e^{\theta_i})^2}$.

3. Experiment Set Up

3.1 Dataset

In this study, we use the Credit Scoring data (Give Me Some Credit). It is available at kaggle.com (https://www.kaggle.com/c/GiveMeSomeCredit). Credit scoring data is used to predict debtors who deserve credit and do not. In this data is available historical data about 150,000 debtors. Of the 150,000 observations, there are some missing values that must be eliminated and used in this study as many as 110290 observations.

3.2 Variable

The dependent variable is $y$. It has a binary value of either 1 or 0. A value of 1 means that the borrower is delinquent and has defaulted on his loans for the last 2 years, while a value of 0 means that the borrower is a good customer and repays his debts on time for the last two years. There are 10 independent variables. Variable $(x_1)$ represents the percentage of total balances in credit cards and private lines of credit except for real estate (no installment debt such as Auto loans) divided by the number of credit limits, variable $(x_2)$ represents the age of debtor, variable $(x_3)$ represents how many times debtor has been 30-59 days late pay (not too bad) in the last 2 years, variable $(x_4)$ represents the percentage of monthly debt payment, allowance, living expenses divided by monthly gross income, variable $(x_5)$ represents the debtor income every month, variable $(x_6)$ represents the number of credit lines and loans, variable $(x_7)$ represents how many times the debtor has been late for pay 90 days or more, variable $(x_8)$ represents denotes the number of mortgage and real estate loans including home equity lines of credit, variable $(x_9)$ represents how many times debtor has been 60-89 days late pay (not too bad) in the last 2 years, and variable $(x_{10})$ represents the number of family members.

4. Result and Discussion

The first step, we do ‘Split’ procedure that is to divide data in the $Q$ data block. We choose $Q = 10$ and then find for $\lambda$ tuning parameter value through a k-cross-validation algorithm. Before doing k-cross-validation process, we must find the value of initial value ($\lambda$) through pathways solution. Selected tuning parameter in each data block will be shown in the following table

| Data block $(q)$ | $\hat{\lambda}$ |
|------------------|------------------|
| 1                | 0.01329002       |
| 2                | 0.001475477      |
| 3                | 0.01944356       |
| 4                | 0.001367119      |
| 5                | 0.002261008      |
| 6                | 0.01088266       |
| 7                | 0.005072335      |
| 8                | 0.001735969      |
| 9                | 0.02407742       |
| 10               | 0.01441134       |
We estimate the parameters by the maximum likelihood estimation (MLE) method and the optimization use the coordinate descent. The procedure of MLE is to maximize the log-likelihood function in equation (7). After maximizing the log-likelihood, we optimize the log likelihood with the coordinate descent. We obtain the parameter estimation value for each data block, then the estimation result from Q data block is combined in final estimation use Conquer method in equation (14) 

**Table 2. Combined estimator**

| Variable | $\hat{B}_j$ |
|----------|-------------|
| Constant | -3.58203    |
| $x_1$    | 1.47142     |
| $x_2$    | -0.00368    |
| $x_3$    | 0.372446    |
| $x_4$    | 0           |
| $x_5$    | 0           |
| $x_6$    | 0           |
| $x_7$    | 0.612527    |
| $x_8$    | 0           |
| $x_9$    | 0.508872    |
| $x_{10}$ | 0           |

Based on the table, we can see that some coefficients are zero. It is due to the LASSO penalty function on penalized logistic regression. LASSO shrinks coefficients on irrelevant variables. So that the irrelevant variables have a small effect on the response variable. The shrinkage of the LASSO coefficient results in a simpler regression model. The relevant variables that exist in the model are $x_1$, $x_2$, $x_3$, $x_7$ and $x_9$

5. Conclusion
The relevant variables present in the model are ($x_1$) represents the percentage of total balances in credit cards and private lines of credit except real estate divided by the number of credit limits, ($x_2$) represents the age of debtor, ($x_3$) represents how many times debtor has been 30-59 days late pay in the last 2 years, ($x_7$) represents how many times the debtor has been late for pay 90 days or more, and ($x_9$) represents how many times debtor has been 60-89 days late pay in the last 2 years.

6. References
[1] Meng C, Wang Y, Zhang X, Mandal A, and Ma P 2017 Effective Statistical Methods for Big Data Analytics pp. 1–22.
[2] Márquez J 2008 An Introduction to Credit Scoring For Small and Medium Size Enterprises,
[3] Chung Y, Shopia R, Vincent D, Andrew G and Jingchen L N 2013 A Nondegenerate Penalized Likelihood Estimator For Variance Parameters In Multilevel Models 78 pp. 685–709.
[4] Tibshirani R 1996 Regression Shrinkage and Selection via the LASSO 5.
[5] Efron B 2004 The Estimation of Prediction Error : Covariance Penalties and Cross-Validation 99 pp. 619–632.
[6] Fonti V and Belitser E 2017 Feature Selection using LASSO.
[7] Gunes F 2015 Penalized Regression Methods for Linear Models in SAS/STAT © SAS Institute Inc pp. 114.
[8] Algamal Z Y and Lee M H 2015 Applying penalized binary logistic regression with correlation based elastic net for variables selection J. Mod. Appl. Stat. Methods.14(1) 168-179.
[9] Krona E 2017 A simulation study of model fitting to high dimensional data using penalized logistic regression Bachelor Thesis Matematiska institution Mathematical Statistics Stockholm University
[10] Yang Y and Zou H 2014 A coordinate majorization descent algorithm for ℓ1 penalized learning J. Stat. Comput. Simul. 84 pp. 84–95.
[11] Friedman J, Hastie T, and Tibshirani R 2010 Regularization Paths for Generalized Linear Models via Coordinate Descent *J. Stat. Softw.* 33(1) pp. 1–24.

[12] Simon N, Friedman J, Hastie T, and Tibshirani R 2011 Regularization Paths for Cox’s Proportional Hazards Model via Coordinate Descent *J. Stat. Softw.* 39(5) pp. 1-13.

[13] Chen X and Xie M 2014 A Split-and-Conquer Approach for Analysis of Extraordinarily Large Data *Statistica Sinica* 24 pp. 1655–1684.