On the selection of regression model using machine learning

Asanao Shimokawa* Etsuo Miyaoka†

Abstract

In this study, we focus on the model selection problem in the logistic regression model. When construct the model from a data set, we need to decide which covariates should be included in the model. In addition to this, it is also need to consider interactions between covariates and their nonlinear transformations. In order to perform these operations automatically, we examine the automatic model selection method using a genetic algorithm. We propose the algorithm for this purpose and verify its performance through simulation studies.

Key Words: Genetic algorithm, Logistic regression, Model selection

1. Introduction

When construct the regression model from a data set, in general, we need to select the set of covariates that must be included in the model. In addition to this, it is also need to determine the presence or absence of interactions between covariates. Conventionally, these tasks are perform by using some criterion such as AIC or BIC. In addition to these tasks, if we want to perform more complete analysis, then we need to investigate the possibility of the nonlinear transformation of the covariates included in the model. These tasks become troublesome when there are many covariates included in the data, and sometimes it is impossible to determine an optimal model using the conventional method. To address this problem, we propose the automatically selection method of a model using machine learning.

In this study, we use the genetic algorithm (GA) for this purpose. The GA refers to a model introduced and investigated by Holland (1975). It is a computational model inspired by evolution. This algorithm encodes a potential solution to a specific problem on a simple chromosome-like data structure and applies recombination operators to these structures to preserve critical information. Although the GA is often viewed as a function optimizer, the range of problems to which it has been applied is quite broad. Siedlecki and Sklansky (1989) showed that the GA is a powerful tool for feature selection when the number of initial feature sets is large. The study showed that the GA method is a powerful tool for large-scale feature selection.

In the next section, we define the notation and model. Further, the details of the proposed model selection method is described. We verify the proposed method through simulation studies. The simulation methods and results are shown in Section 3. In our simulation studies, we especially focus on the model selection problem for logistic regression model. Finally, we conclude this paper in Section 4.

2. Method

2.1 Notation and Model

Let Y and $\boldsymbol{x} = (x_1, x_2, \cdots, x_p)$ be the outcome and independent variaboles, respectively. In this study, we assume that Y takes two values 1 or 0. An observed data is represented

*Tokyo University of Science, 1-3 Kagurazaka, Shinjuku-ku, Tokyo 162-8601, Japan
†Tokyo University of Science, 1-3 Kagurazaka, Shinjuku-ku, Tokyo 162-8601, Japan

by $\mathscr{L} = \{(\boldsymbol{x}_i, y_i); i = 1, 2, \cdots, n\}$. Let the conditional probability that the outcome is present be denoted by $Pr(Y = 1|\boldsymbol{x}) = \pi(\boldsymbol{x})$. Then, as in the logistic regression case, we want to construct the logit model which is given by the equation

$$g(\boldsymbol{x}) = \ln\left(\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}\right) = \beta_1 h_1(\boldsymbol{x}) + \beta_2 h_2(\boldsymbol{x}) + \cdots + \beta_q h_q(\boldsymbol{x}),$$

where $h_1(\cdot), h_2(\cdot), \cdots, h_q(\cdot)$ are some functions. For example of $h_k(\boldsymbol{x})$, $h_k(\boldsymbol{x}) = x_j^r$, $h_k(\boldsymbol{x}) = \log(x_j)$, and $h_k(\boldsymbol{x}) = x_{j_1}x_{j_2}$, $(j, j_1, j_2 = 1, 2, \cdots, p, k = 1, 2, \cdots, q, r \in \mathcal{R})$.

As the function of $h_k(\cdot)$ for our simulation studies, we use the power functions used in Royston and Altman (1994):

$$h_k(\boldsymbol{x}) = \begin{cases} x_j^{r_k}, & r_k \neq r_{k-1} \\ h_{k-1}(\boldsymbol{x})\ln(x_j), & r_k = r_{k-1} \end{cases},$$

where $r_k \in \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$ and $r_k = 0$ denotes the log of the variable, and interactions of two variables:

$$h_k(\boldsymbol{x}) = x_{j_1} \times x_{j_2}, \quad j_1 \neq j_2.$$

2.2 Model Selection Using GA

In general logistic regression analysis, the stepwise selection method or some other methods are used to select the best model from many variables and combinations of their transformations. However, this task is very troublesome, and sometimes it is impossible to determine the optimal model. Therefore, we consider the model selection by GA.

The first step in the implementation of the GA is to generate an initial population. Each member of this population will be represented as a binary string of length L, which corresponds to the problem encoded. Each string is referred to as a "chromosome." In this case, each chromosome represents one logit model (For example $g(\boldsymbol{x}) = \beta_1 + \beta_2 x_1 + \beta_3 \log(x_1) + \beta_4 x_2 x_3$).

The execution of the GA can expressed as a two-stage process. In the first stage, it starts with the current population, and selection is applied to create an intermediate population. The selection is carried out on the basis of the evaluation f_l for each chromosome a_l. As the evaluation function, we used the AIC, BIC, and prediction error in this study.

In the second stage, crossover and mutation are applied to the intermediate population to create the next population. The process of proceeding from the current population to the next population is one generation in the GA. After the process of selection, crossover, and mutation is complete, the next population will be evaluated. The process of evaluation, selection, recombination, and mutation corresponds to one generation in the execution of a GA.

An algorithmic description of the GA used in this study is given below:

1. Generate the initial population randomly for the chromosomes a_l.
2. The current population: $\Pi = \{a_l\}, l = 1, 2, \cdots, m$.
3. **For** $g \leftarrow 1$ **to** the number of generations G **do**
4. Initialize the intermediate population I and the offspring population O.
5. **For** $l \leftarrow 1$ **to** the number of chromosomes m **do**
6. Evaluate the chromosome a_i in the population Π.

Table 1: Mean and standard deviation of the number of coefficients selected by GA in our simulataions.

Model	Mean of # coefficients selected by GA (std.)
Model 1	7.51 (1.93)
Model 2	7.04 (1.98)
Model 3	7.75 (1.50)
Model 4	7.89 (1.98)

7. Copy to the intermediate population I from Π on the basis of the evaluation of chromosomes
   ($I = \{a'_l\}, l = 1, 2, \cdots, m$).
8. **For** $k \leftarrow 1$ **to** $n/2$ **do**
9. Choose the two parents a'_{l_1} and a'_{l_2} at random from M, and apply performed with the probability P_c
   (a''_{l_1},$a''_{l_2} = (a'_{l_1}, a'_{l_2}) \cup$ crossover(a'_{l_1}, a'_{l_2})).
10. Add a''_{l_1} and a''_{l_2} to the offspring population O.
11. The offspring population: $O = \{a''_l\}, l = 1, 2, \cdots, m$.
12. **For** $i \leftarrow l$ **to** the number of chromosomes m **do**
13. **For** $b \leftarrow 1$ **to** the number of bits L **do**
14. Apply with mutation probability P_m to the bth bit from the chromosome a''_l in O.
15. Replace population Π by the offspring population O.
16. Evaluate each chromosome in population Π of the last generation and get the best chromosome.

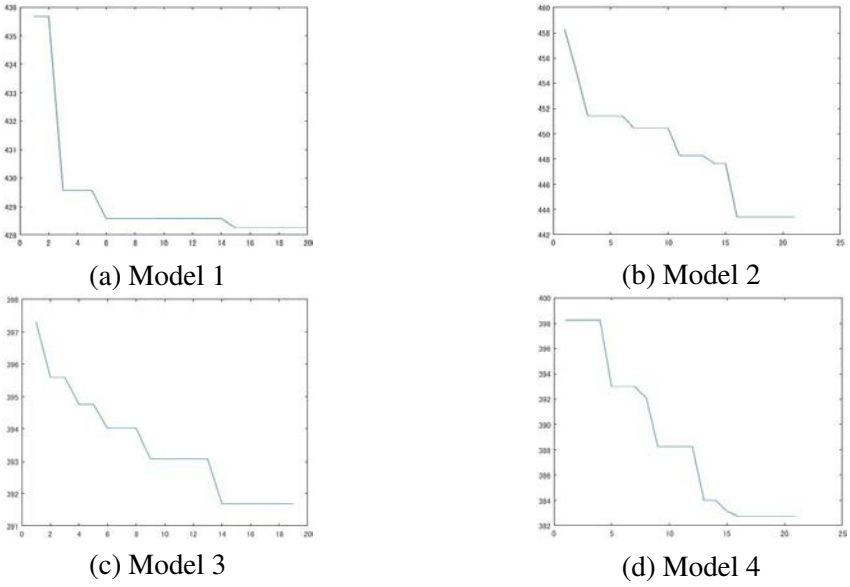
3. Simulation

We present simple simulation studies to study the method described in previous section. The covariates x_j are generated from the uniform distributions between 0 and 1 ($j = 1, 2, \cdots, 5$). We used the data generated from four models:

$$
\begin{aligned}
Model1: &\quad g(\boldsymbol{x}) = 1 + x_1 \\
Model2: &\quad g(\boldsymbol{x}) = 1 + x_1^2 \\
Model3: &\quad g(\boldsymbol{x}) = 1 + \log(x_1) \\
Model4: &\quad g(\boldsymbol{x}) = 1 + x_1^2 + \log x_1 + \log(x_2) + x_1 \times x_2.
\end{aligned}
$$

The number of samples is set to $n = 300$, and simulations are repeated 100 times.

Table 1 shows the mean and standard deviation of the number of coefficients selected by GA. Figure 1 shows the typical plots of the evaluations of the chromosome that are selected in each generations in GA. As the results of these simulations, GA work reasonably well in selecting variables within a logistic model. On the other hand, there is a tendency to include many coefficients in the model. This tendency will continue even if the number of generations is increased.



(a) Model 1 (b) Model 2

(c) Model 3 (d) Model 4

Figure 1: The typical plots of the evaluations of the chromosome that are selected in each generations in GA. The horizontal axis represents the generations. The vertical axis represents the evaluations (AIC) of the selected chromosome in each generations.

4. Conclusion

In this study, we focus on the model selection problem in the logistic regression model. We research the automatic model selection method using a genetic algorithm. We propose the algorithm for this purpose and verify its performance through simulation studies. As the main results obtained by simulations studies, GA work reasonably well in selecting variables within a logistic model. Future challenges include dealing with the tendency to include many coefficients in the model.

REFERENCES

Holland J. H. (1975), *Adaptation in Natural and Artifical Systems*, University of Michigan Press.

Siedecki W. and Sklansky J. (1989) "A note on genetic algorithms for large-scale feature selection," *Pattern Recogn. Lett.*, 10(5), 335–347.

Royston P. and Altman D. G. (1994), "Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling (with discussion)", *Applied Statistics*, 43, 429 – 467.

Akaike, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle," in *Second International Symposium on Information Theory*, eds. B. N. Petrov and F. Càski, Budapest: Akademiai Kiàdo, 267–281.