Journal of Physics: Conference Series

PAPER • OPEN ACCESS

Parameter estimation for high dimensional classification model on colon cancer microarray dataset

To cite this article: Bahriddin Abapihi *et al* 2021 *J. Phys.: Conf. Ser.* **1899** 012113

View the article online for updates and enhancements.

Parameter estimation for high dimensional classification model on colon cancer microarray dataset

Bahriddin Abapihi[1], Mukhsar[1*], Gusti Ngurah Adhi Wibawa[1], Baharuddin[1], Favorisen Rosyking Lumbanraja[2], Mohammad Reza Faisal[3] and Asrul Sani[4]

[1]Department of Statistics, Halu Oleo University, Kendari, Indonesia,
[2]Department of Computer Science, Lampung University, Bandar Lampung, Indonesia,
[3]Department of Computer Science, Lambung Mangkurat University, Banjarmasin, Indonesia,
[4]Department of Mathematics, Halu Oleo University, Kendari, Indonesia

mukhsarmuhu@gmail.com (corresponding author)

Abstract. In classification problems, logistic regression is among powerful techniques for discrimination. It provides directive probabilities of sample classification and interpretable coefficients. When it comes to model high dimensional dataset, however, logistic regression with its Newton-Raphson method of parameter estimation is no longer applicable, especially on low sample size and extremely high dimension. By applying cross-entropy algorithm on regularized logistic regression, it was able to well performing parameter estimation and highly accurate classification result.

1. Introduction

The advanced technologies in molecular biology and genetics such as DNA microarrays lead to the whole genome analysis instead of the study of genes individually. In a single experiment using the DNA microarray technology, the levels of thousands of gene expression can be measured. Biomedical research is one of the fields that take benefits from this technology. There are so many cancer classification based datasets resulted from microarray gene expression technology. With these datasets it allows us to compare the levels of gene expression on different instances in samples (tissues).

The problem of small sample size and high dimensionality in microarray dataset is still a challenging issue in cancer classification. The dataset is typically to have number of features (genes) more than the number of samples (tissues). Usually, the number of samples is about a hundred or less, while the number features is more than one thousand. Of these thousands features, most of them are irrelevant and only a small number of important genes are really responsible to cancer classification. Due to these characteristics of the datasets, the development of techniques for the purpose of feature selection remains a challenging task. A good selection of features could improve the classification accuracy and could also be providing a much simpler interpretation of the result [1, 2, 3].

Many authors have been proposing their methods for classification and feature selection almost any field of applications [2, 3]. A well known classical and powerful method for classification is logistic regression. This technique provides predictive probability of classification and easy interpretation of coefficients. However, when it comes to model high dimensional dataset which is also to have low sample size, the Newton-Raphson method of parameter estimation for logistic regression model is no longer applicable. To overcome this situation, we proposed cross-entropy algorithm for parameter estimation on regularized logistic regression.

2. Regularized Logistic Regression

Logistic regression is a regression model to be used for binary classification [10]. Usually, the response variable has two values either 1 or 0. In this model the linear combination of predictor variables has a nonlinear relation with the binary responses. Suppose that we have n samples and p variables. Let the response variable denoted by $y_i \in [0, 1]$ and $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ be the vector of predictor variables. The relation between response and predictor variables is formulated by

$$\pi_i = p(y_i = 1|\mathbf{x}_i) = \frac{\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})}{1+\exp(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})}, \qquad i = 1, 2, \ldots, n \tag{1}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is a vector of p unknown coefficients of predictor variables. For simplicity in estimation, we take the log-likelihood function (logit transformation) of Equation (1) as

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n}[y_i \log(\pi_i) + (1 - y_i)\log(1 - \pi_i)]\,. \tag{2}$$

To control the size of variable coefficients, we can add a regulation term to the log-likelihood function. Tibshirani [4] proposed l_1-norm regularization to simultaneously conduct variable selection and estimation. This regularization is also known as the least absolute shrinkage and selection operator (LASSO). Note that the log-likelihood function, $l(\boldsymbol{\beta})$, is negative, while the regularization term is nonnegative. Consequently, estimation is conducted by adding the constraint to the negative log-likelihood function. As a result, the regularized logistic regression using LASSO is defined as

$$\hat{\boldsymbol{\beta}}_{LASSO} = \operatorname{argmin}_{\boldsymbol{\beta}}\left[-\sum_{i=1}^{n} y_i \log(\pi_i) + (1 - y_i)\log(1 - \pi_i) + \lambda\sum_{j=1}^{p}\left|\beta_j\right|\right] \tag{3}$$

The constant λ is a tuning parameter. When $\lambda = 0$, it reduces to maximum likelihood estimator, while $\lambda \to \infty$, it forces all coefficients of variables to be zeros [4].

3. The Cross-Entropy (CE) Algorithm

The CE algorithm provides a simple and efficient method to solve optimization problems. Since most of estimations are in relation to optimization (maximization or minimization) the CE can deal with this problem. The reputation of CE has been proven in almost any field of application, such as reliability analysis, performance analysis of telecommunication problem, and operation research [6, 7, 8, 9].

The CE involves two processes iteratively, random sample generation and parameter update of random mechanism in order to produce a better sample in the next iteration. In general, the CE algorithm for estimation problem is outlined as follows [7, 8].

- Generate random samples based on a mechanism (for example using normal distribution) as the candidates of parameter to estimate
- Fit these random samples into the objective function
- Sort the samples based on fitting values of objective function
- Take some elite samples (the samples with best fitting) and calculate the mean and the variance
- Repeat step 1 to 4 until the stopping criteria satisfied.

4. The Proposed Method and Dataset

In this chapter we are going to explain the procedure of our method. The algorithm of our method is organized as follows.

- Applying leave-one-out cross-validation (LOOCV) in this research, we separated the samples into training and testing. One sample was for testing and the rests were for training.
- After separating samples, we fixed the training samples into LASSO logistic regression model and estimated the parameters (coefficients) using the cross-entropy algorithm. These

coefficients were, then, sorted in related to their absolute values from the largest to the smallest.

- Based on these coefficients, we conducted support vector machine (SVM) classification starting from variable which has the largest coefficient, and calculated the predicted accuracy. By adding one by one the second variable of the second largest coefficient until adding variable of the smallest coefficient, the SVM classification took place and the predicted accuracies were calculated.
- We plotted the accuracy trajectory of variables ordered by their absolute coefficients. We selected combination of variables that produced the highest accuracy.
- Using selected variables, we repeated step 2 to 4 until the maximum accuracy be reached.

The dataset used in this paper is colon cancer dataset. It is freely available from the web as it also published by [5]. This dataset consists of 62 samples and 2000 variables (genes). Of the 62 samples, 40 samples are cancerous and 22 others are noncancerous,

5. Result and Discussion

The experiment was conducted in 6 iterations. In the first iteration we selected 148 of 2000 variables. The trajectory plot can be seen in Figure 1.

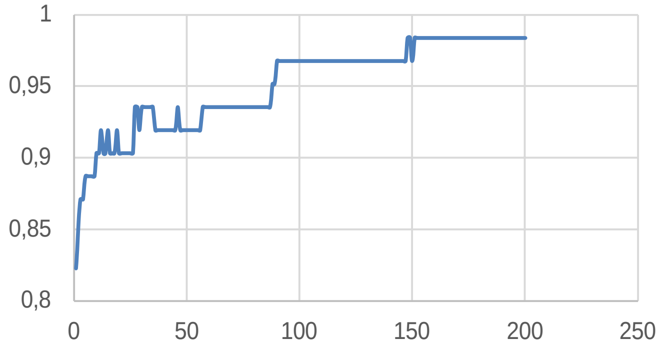

Figure 1. Accuracy trajectory plot in the first iteration with vertical axis for accuracy and horizontal axis for number of selected variables

Table 1. The accuracy and number of selected variables in each iteration

Iteration	1	2	3	4	**5**	6
Number of variables before parameter estimation	2000	148	76	51	**28**	26
Accuracy using all variables after parameter estimation	0.8065	0.9194	0.9355	0.9677	**0.9839**	0.9677
Number of selected variables	148	76	51	28	**26**	26
Accuracy after sorting and selecting variables	0.9839	1.0000	1.0000	1.0000	**1.0000**	0.9839

From Table 1, we can see that the best performance was reached in iteration 5 with the number of selected variables is 26 and the accuracy is 1.0000.

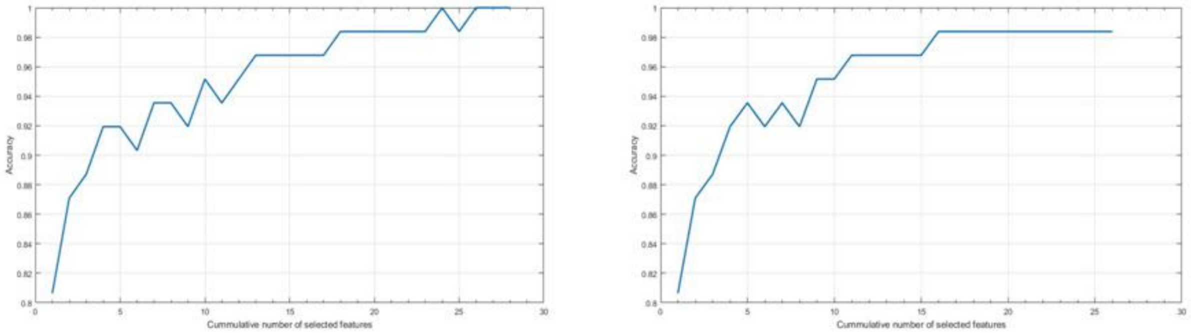

Figure 2. Accuracy trajectory plot in iteration 5 (left) and iteration 6 (right)

As seen in Figure 2, the best accuracy was achieved in iteration 5. Accordingly, the selection of variables should be based on the result in iteration 5. The coefficient value of each variable can be seen in Table 2. Note that the accuracy in Table 2 is a cumulative accuracy of the top best variables. In other words, Table 2 also shows the accuracy, when a model is using variable of first rank, is 0.807, and then by adding one variable of the next rank, it is 0.871, and so on. The perfect accuracy, 1.0, is achieved when using up to 26 variables.

Table 2. List of variables with their ranks, coefficient values and cumulative accuracies

Table 2a. Rank 1 to 13

Rank	1	2	3	4	5	6	7	8	9	10	11	12	13
Variable No.	765	377	1772	1241	353	14	1976	493	792	1769	1482	554	1644
Coefficient	-1.52	-1.14	0.74	0.64	0.55	-0.45	-0.41	-0.35	-0.32	0.29	-0.26	-0.25	-0.25
Accuracy	0.807	0.871	0.887	0.919	0.919	0.903	0.936	0.936	0.919	0.952	0.93	0.952	0.96

Table 2b. Rank 14 to 26

Rank	14	15	16	17	18	19	20	21	22	23	24	25	26
Variable No.	1423	974	70	1325	783	590	1740	1560	164	1641	1346	1058	995
Coefficient	-0.21	0.21	-0.21	0.20	0.20	0.17	0.17	-0.17	0.15	0.12	0.11	-0.07	0.04
Accuracy	0.968	0.968	0.968	0.968	0.984	0.984	0.984	0.984	0.984	0.984	1	0.984	1

6. Conclusion

We have shown that by applying the cross-entropy algorithm on regularized logistic regression high performance accuracy can be achieved. Our proposed method can be applied on dataset that has high dimensionality and low sample size.

References

[1] Kalina, J. 2014. Classification methods for high dimensional genetic data. Biocybern, Biomed. Eng. (34) 10 – 18.

[2] Ma, S. & Huang, J. 2008. Penalized feature selection and classification in bioinformatics. Brief. Bioinform. (9) 392 – 403.

[3] Chandra, B. & Gupta, M. 2011. An efficient statistical feature selection approach for classification of gene expression data. J. Biomed. Inform. (44) 529 – 535.

[4] Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B (Methodological) (58) 267 – 288.

[5] Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D. & Lavine, A. J. 1999. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Natl. Acad. Sci. (96) 6745 – 6750.

[6] Rubinstein, R.Y., 1997. Optimization of computer simulation models with rare events. Eur. J. Oper. Res. 990 (1), 89–112.

[7] Rubinstein, R.Y., 1999. The cross-entropy method for combinatorial and continuous optimization. Methodol. Comput. Appl. Probab. 10 (2), 127–190.

[8] Rubinstein, R.Y., 2001. Combinatorial optimization, cross-entropy, ants and rare events. In: Uryasev, S., Pardalos, P.M. (Eds.), Stochastic Optimization: Algorithms and Applications, Kluwer, Dordrecht, pp. 304–358.

[9] Rubinstein, R.Y., Kroese, D.P., 2004. The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization. Monte Carlo Simulation and Machine Learning. Springer-Verlag, New York

[10] Collet, D. 2002. Modelling Binary Data. New York, Chapman-Hall.