Proceedings of Science and Mathematics
Faculty of Science, Universiti Teknologi Malaysia

Vol. 7, 2022, Page 38 - 42

Correlation Filtered Sure Independence Screening: Multinomial Logistic Regression

Demudu Naganaidu[a, b*], Zarina Mohd Khalid[b]
[a]Centre for Postgraduate Studies, Asia Metropolitan University
[b]Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia
*Corresponding author: demudu@amu.edu.my

Abstract

Contemporary variable selection via penalization or regularization for high-dimensional is data break down when data grows into ultrahigh dimensional space. Commonly used methods to overcome this problem is by screening before applying the penalization. Sure independence screening (SIS) selects variables based on the marginal utility ranking before regularization is applied on the model. However marginal utility from highly correlated variables tends to carry redundant information. Hence, in this paper, 3 stage variable selection method is implemented. Stage 1 is to select ***n*** -1 variables from ranked marginal utility. Stage 2 is to use a simple bivariate correlation to remove the highly positively correlated variables from the n-1 variables. Finally, Stage 3 is to apply the Least Absolute Shrinkage and Selection Operator (LASSO) regularization technique to further select variables automatically. Eight alternative models developed to evaluate the proposed method and tested on microarray gene expression dataset series GSE65194 for breast cancer type classification. The correlation-filtered model still produced high accuracy, but with fewer variables or genes in cancer classification. Over-pruning of variables, on the other hand, causes model accuracy to deteriorate.

Keywords: Ultrahigh-dimensional data; Sure Independence Screening; LASSO; Multinomial Logistic Regression; Variable Selection

Introduction

High-dimensional data is a case where the number of variables, p , exceed the number of observations, n, frequently termed as the “large p and small n problem” ($p \gg n$) [1, 2]. There are several challenges encountered in high-dimensional data, such as data visualizations, computational complexity [3, 4], overfitting [5] and poor interpretability due to a large number of variables. Variable selection via regularization is often the researchers’ option solving the problem of $p \gg n$ [6].

Though a regularization method such as Least Absolute Shrinkage and Selection Operator (LASSO) [7] is a promising method in automatic variable selection in high-dimensional data, this method breaks down when data is in ultrahigh dimensional space. Ultrahigh-dimensional data refers to a dataset with $\log(p) = O(n^{\alpha})$ for some $0< \alpha<1$ [8]. Selecting variables in ultrahigh-dimensional models correctly and automatically is a tough problem. Fan and Lv [8] introduced a breakthrough method known as sure independence screening (SIS) to address this problem in regression context. All-important variables in the model are retained with a probability close to 1 when variables are selected using the SIS method. The method is simple, straightforward, and computationally efficient, with the goal of lowering the number of independent variables before variable selection via regularized model learning, often known as two stage variable selection.

In this era of modern technologies with remarkable computing facilities, data is collected often in ultrahigh dimensional space. For example, microarrays, financial data, astronomical data, and image processing data. In gene expression data for cancer classification, it is common to have tens of thousands of variables while the number of observations or samples are only in tens or hundreds [9]. However, only a small number of genes are relevant to the disease while other genes are just noise [10, 11]. In addition, many genes are highly correlated, resulting in redundant data. Effective gene selection is important in predicting cancer types with high prediction accuracy.

Multinomial Logistic Regression (MLR) is part of the generalized linear model (GLM) family suitable for multiclass classification problems [10, 12]. SIS method was extended to GLM by Fan et al. [9]. The negative log-likelihood for each independent variable is defined as marginal utility. The marginal utilities are ranked in ascending order and $\boldsymbol{q}$-vector independent variables selected, usually $\boldsymbol{q} < \boldsymbol{n}$, before applying regularized model learning. However, the first $\boldsymbol{q}$ variables selected could be highly correlated variables carrying redundant information, resulting in a complex model with possible overfitting. To address this problem, 3 stages variable selection is proposed. Stage 1 – Variable selection via SIS. Stage 2 – Variable selection via correlation filtering on variables selected in Stage 1. Stage 3 – Variable selection via regularization on variables selected in Stage 2. A correlation filtering at Stage 2 removes variables which are highly correlated, resulting in a more parsimony model for better interpretation and reducing overfitting.

Materials and methods

Let dataset with $\boldsymbol{K}$ category response variable $\boldsymbol{Y_i} \in [\boldsymbol{0}, \boldsymbol{1}, .., \boldsymbol{K}]$ and $\boldsymbol{X} = [\boldsymbol{X_0}, \boldsymbol{X_1}, .., \boldsymbol{X_p}]$ where $\boldsymbol{X_0} \equiv \boldsymbol{1}$ be independent variables that influence the response variable. $\boldsymbol{Y_i} \sim \boldsymbol{multinomial}\,(\boldsymbol{n} = \boldsymbol{1}, \boldsymbol{p} = (\boldsymbol{p_{i1}}, .., \boldsymbol{p_{iK}}))$ subject to:

$$\sum_{i=1}^{K} \quad \boldsymbol{Y_{ik}} = \boldsymbol{1} \text{ and } \sum_{i=1}^{K} \quad \boldsymbol{p_{ik}} = \boldsymbol{1} \qquad (1)$$

Following Hosmer and Lameshow [13] the MLR model in the logit form taking category $\boldsymbol{0}$ as the reference category can written as following equation:

$$g_j(x) = ln\left\{\frac{P(Y = j|x)}{P(Y = 0|x)}\right\} = \beta_{j0} + \beta_{j1}x_1 + , .., \qquad \beta_{jp}x_p \;, j = 1,2, .., K \qquad (2)$$

where $\boldsymbol{g_j}(\boldsymbol{x})$ is the logit function, $\boldsymbol{\pi_j}(\boldsymbol{x}) = \boldsymbol{P}(\boldsymbol{x}) = \frac{e^{\boldsymbol{g_j}(\boldsymbol{x})}}{\sum_{k=0}^{K} \quad \boldsymbol{g_j}(\boldsymbol{x})}$, j = 0,1,..,K and $\boldsymbol{\beta_{jp}}$ is the coefficient for logit function $\boldsymbol{g_j}(\boldsymbol{x})$ for independent variable $\boldsymbol{p}$ can be estimated via maximum likelihood $\boldsymbol{L}(\boldsymbol{\beta})$ or equally log-likelihood $\boldsymbol{l}(\boldsymbol{\beta})$ respectively written as follows.

$$\boldsymbol{L}(\boldsymbol{\beta}) = \prod_{i=1}^{n} \quad \prod_{j=0}^{K} \quad [\,\boldsymbol{\pi}_{ij}^{Y_{ij}}(\boldsymbol{x})] \text{ and } \qquad (3)$$

$$\boldsymbol{l}(\boldsymbol{\beta}) = \sum_{i}^{n} \quad [\boldsymbol{y_{1i}}\boldsymbol{g_j}(\boldsymbol{x_i}) + \cdots + \boldsymbol{y_{ji}}\boldsymbol{g_j}(\boldsymbol{x_i}) - \boldsymbol{ln}\,(\boldsymbol{1} + \boldsymbol{e}^{g_1(x_i)} + .. + \boldsymbol{e}^{g_j(x_i)}\,] \qquad (4)$$

Having these formulated, the 3 stages variable selection are performed as follows:

Stage 1: Variable Selection via SIS

All independent variables are standardized via a robust standardization method to minimize the effect of different scales of measurements. Following the SIS method [8], the marginal utility $\boldsymbol{L_j}$ which is the negative log-likelihood are computed. The marginal utility for the $\boldsymbol{j}$th independent variable $\boldsymbol{x_j}$ for $\boldsymbol{j} = \boldsymbol{1}, .., \boldsymbol{p}$, with response variable $\boldsymbol{y_i},\ \boldsymbol{i} = \boldsymbol{1}, .., \boldsymbol{n}$, is defined by:

$$\boldsymbol{L_0} = \boldsymbol{l}(\boldsymbol{\beta}) = \boldsymbol{log\ log\ L}(\boldsymbol{y_i}, \boldsymbol{\beta_0}) \text{ and } \boldsymbol{L_j} = \boldsymbol{l}(\boldsymbol{\beta}) = \boldsymbol{log\ log\ L}(\boldsymbol{y_i}, \boldsymbol{\beta_0} + \boldsymbol{x_{i,j}}\boldsymbol{\beta_j}) \qquad (5)$$

The marginal utilities $\boldsymbol{L_1,..,L_p}$ are then ranked in ascending order giving, $\boldsymbol{L_{v(1)}, L_{v(2)},.., L_{v(q)},..., L_{v(p)}}$ from where $\boldsymbol{q}$ vector of independent variables $(\boldsymbol{x_{v(1)}, x_{v(2)},.., x_{v(q)}})$ selected for Stage 2 screening. Here, $\boldsymbol{q = n - 1}$, as suggested by Fan and Lv [8]. With, $\boldsymbol{q < n}$ **,** computational complexity is reduced, and low dimensional statistical methods can be applied.

Stage 2: Variable Selection via Correlation Filtering

The $\boldsymbol{q}$ variables selected from Stage 1, further filtered with a simple bivariate correlation, $\boldsymbol{r_{jl}}$, between independent variable $\boldsymbol{j}$ and $\boldsymbol{l}$**,** computed by the following formula:

$$r_{jl} = \frac{\sum_{i=1}^{n} \quad (x_{ij} - \underline{x}_j)(x_{il} - \underline{x}_l)}{\sqrt{\sum_{i}^{n} \quad (x_{ij} - \underline{x}_j)^2 (x_{il} - \underline{x}_l)^2}} \tag{6}$$

When two variables are correlated with a cut-off correlation greater than or equal to $\boldsymbol{r}$, one of the variables will be dropped. A positive correlation indicates both carry similar information, thus impacting the response variable in the same way; hence one variable is redundant and can be dropped. The cut-off correlation is selected to be ***0.7***, ***0.8*** or ***0.9*** in this paper. An only positive correlation is considered as the negative correlation indicates the opposite impact of the independent variable to the response variable. Denoting the variables selected as $\boldsymbol{m}$ after positive correlation filtering, where $\boldsymbol{m} \leq \boldsymbol{q}$,

Stage 3: Variable Selection via Regularization

The ***m***, variables selected in Stage 2, may still include many unimportant independent variables. This is addressed via LASSO [7] regularized likelihood in Stage 3. The final parameter estimation is defined as:

$$\widehat{\boldsymbol{\beta}} = \left[\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \quad \sum_{\boldsymbol{l=0}}^{\boldsymbol{m}} \quad \boldsymbol{l(\beta)} + \boldsymbol{\lambda} \sum_{\boldsymbol{k=0}}^{\boldsymbol{K}} \quad \sum_{\boldsymbol{l=1}}^{\boldsymbol{m}} \quad \boldsymbol{\beta_{kl}} \right] \tag{7}$$

where λ is the tuning parameter selected via cross validation. The final estimated parameter matrix is sparse, where non-zero columns, ***c* << *n***.

To the best knowledge the additional filtering of variables via correlation after sure independence screening has not been implemented yet in literature.

To demonstrate the usefulness of the suggested methodology, 8 alternative models developed and tested with a real data set. Four models consist of MLR models and four MLR models with LASSO regularization (LASSO MLR).

Microarray dataset from breast cancer samples with the identification GSE65194 were downloaded from the Gene Expression Omnibus (GEO) of the National Center for Biotechnology Information (NCBI). This dataset consists of only 178 observations but with a whopping of 54763 independent variables but with no missing values. The responses variable consists of six classes i.e normal tissue (Healthy) and five different types of cancer known as Luminal A, Luminal B, Triple Negative Breast Cancer/Basal-like (TNBC), TNBC cell lines (Cell line), (Human Epidermal Growth Factor Receptor 2 (Her2). Split ratio of 70:30 applied on data for training set and test set for model building and model testing respectively.

Results and discussion

Model development and evaluation of the proposed method is implemented with Python Programming. Confusion matrix and classification report available in Python Scikit-Learn library [14] are used to compute these performance metrics:

a. Accuracy: Metric to measure model ability to correctly identify each cancer type.

b. Sensitivity: Metric to measure model ability to identify true positive of each cancer type.

c. Specificity: Metric to measure model ability to identify true negative of each cancer type.

Table 1: Sensitivity and specificity values for test dataset

Model trained on 70% of dataset	**Correlation used to filter variables**	**Sensitivity (Specificity)**						**Accuracy On 30% of dataset [10 fold cross validation]**	**Genes Selected**		
		Cancer Type							**Stage**		
		0	**1**	**2**	**3**	**4**	**5**		**1**	**2**	**3**
1. MLR	-	0.75 (1.00)	0.75 (1.00)	0.91 (1.00)	1.00 (0.98)	0.9 (1.00)	0.95 (1.00)	0.95	177	NA	NA
2. MLR	0.90	0.75 (1.00)	0.75 (1.00)	0.91 (1.00)	1.0 (0.98)	0.90 (1.00)	0.95 (1.00)	0.96	177	94	NA
3. MLR	0.80	0.75 (1.00)	0.75 (1.00)	0.91 (1.00)	1.00 (0.98)	0.88 (1.00)	0.95 (1.00)	0.96	177	21	NA
4. MLR	0.70	0.11 (1.00)	0.27 (1.00)	0.71 (0.90)	0.69 (0.90)	0.33 (1.00)	0.70 (1.00)	0.89	177	5	NA
5. LASSO MLR	-	0.75 (1.00)	0.75 (1.00)	0.91 (1.00)	0.89 (1.00)	1.00 (0.98)	0.95 (1.00)	0.82	177	NA	49
6. LASSO MLR	0.90	0.75 (1.00)	0.75 (1.00)	0.91 (1.00)	0.89 (1.00)	1.00 (0.98)	0.95 (1.00)	0.80	177	94	40
7. LASSO MLR	0.80	0.60 (1.00)	0.60 (1.00)	0.82 (1.00)	0.90 (0.98)	0.90 (0.98)	0.90 (1.00)	0.77	177	18	15
8. LASSO MLR	0.70	0.00 (1.00)	0.21 (1.00)	0.67 (0.85)	0.60 (0.87)	0.15 (1.00)	0.63 (1.00)	0.72	177	5	5

Cancer Type: 0 – Healthy, 1- Cell line, 2- Her, 3- Lumina A, 4- Lumina B, 5- TNBC

The results are summarized in Table 1. From Table 1, Stage 1 resulted 177 genes selected to be included in the model. High sensitivity and specificity is recorded for the MLR model (1), but it is not a parsimonious model in relation to other models in Table 1. In model (2), with a cut off correlation of 0.90, the number of genes lowered to 94 with slightly increased accuracy of 0.96. Lower cut-off correlation of 0.80 and 0.70 has further reduced the number of genes respectively to 21 and 5 but with deteriorating sensitivity and specificity performance metrics. LASSO MLR models (5) to (8) generated the final genes with the same trend of MLR models. LASSO MLR model (6) achieved an accuracy of 0.80 compared to LASSO MLR model (5) of 0.82, but the final number of genes is an improvement from 49 to 40.

Conclusion

This paper investigates the problem of ultrahigh dimensional space variable selection for the MLR model. The concept of three-stage variable selection is proposed. In stage 1, the existing SIS method lowers the number of independent variables from high dimensional space to low dimensional space. In stage 2, these variables are further reduced if any two variables are highly correlated by filtering one of the variables. Finally, in stage 3, the regularization method performs automatic variable selection. Correlation filtering in stage 2, able to further reduce additional variables by removing variables that are highly correlated, giving an improved parsimony model. The performance metrics, however, deteriorated when correlation **filtering was less** than 0.80. A more parsimony classification model is essential for researchers to focus on genes responsible for different cancer types. The genes **identified are useful** in predicting cancer type in different subjects and will assist doctors investigate a patient's danger profile and to prescribe a route of treatment tailor-made to that profile.

Acknowledgement

The authors would like to express their appreciation for the support from Universiti Teknologi Malaysia, UTMER2021 Project Number PY/2021/01377.

References

[1] Hastie T et. all. 2009. Springer Series in Statistics The Elements of Statistical Learning. *Math Intell* ; 27: 83–85.

[2] Alfons A, Croux C, Gelper S. 2013. Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *Ann Appl Stat* ; 7: 226–248.

[3] Giraud C. 2015. *Introduction to High-dimensional Statistics*. Epub ahead of print 2015. DOI: 10.1111/insr.12145.

[4] Zhang C, Guo J, Lu J. 2017. Research on Classification Method of High-Dimensional Class-Imbalanced Data Sets Based on SVM. *Proc - 2017 IEEE 2nd Int Conf Data Sci Cyberspace, DSC 2017*; 60–67.

[5] Piao Y, Piao M, Park K, et al. 2012. An ensemble correlation-based gene selection algorithm for cancer classification with gene expression data. *Bioinformatics*; 28: 3306–3315.

[6] Haut JM, Liu Y, Paoletti ME, et al. 2018. Evaluation of different regularization methods for the extreme learning machine applied to hyperspectral images. *Int Geosci Remote Sens Symp*; 2018-July: 3603–3606.

[7] Tibshirani R. 1996. Regression Shrinkage and Selection Via the Lasso. *J R Stat Soc Ser B*; 58: 267–288.

[8] Fan J, Lv J. 2008. Sure independence screening for ultrahigh dimensional feature space. *J R Stat Soc Ser B Stat Methodol*; 70: 849–911.

[9] Fan J, Samworth R, Wu Y. 2009. Ultrahigh Dimensional Feature Selection: Beyond the Linear Model. *Journal of Machine Learning Research* ; 10: 2013–2038.

[10] Kim Y, Kwon S, Heun Song S. 2006. Multiclass sparse logistic regression for classification of multiple cancer types using gene expression data. *Comput Stat Data Anal*; 51: 1643–1655.

[11] Asenso TQ, Zhang H, Liang Y. 2020. Pliable lasso for the multinomial logistic regression. *Commun Stat - Theory Methods*; 0: 1–16.

[12] Krishnapuram B, Carin L, Figueiredo MAT, et al. 2005. Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *IEEE Trans Pattern Anal Mach Intell*; 27: 957–968.

[13] Hosmer DW, Lemeshow S, Sturdivant RX. *Applied Logistic Regression: Third Edition*. 2013. Epub ahead of print 2013. DOI: 10.1002/9781118548387.

[14] Fabian P, Ga¨el V, Alexandre G, et al. 2019. Scikit-learn: Machine Learning in Python Fabian. *Environ Health Perspect*; 127: 2825–2830.