Variable Selection Using Nearest Neighbor Rule in Discriminant Analysis of Dichotomous Data

Kyubark Shim

Abstract:In recent years, as interest in discriminant analysis of big data has increased, research on this field is active. The use of Big Data is also essential for politics and social issues that are sensitive to public opinion. Classifying precisely which respondents belong to which group is very helpful for policy formulation. Kim et al. (2013) propose an intelligent Voice of customer(VOC) analyzing system based on opinion mining to discriminant the unstructured VOC data automatically and determine the polarity as well as the type of VOC. In this paper, based on the previous studies, we selected the variables that can minimize the discrimination error by using the nearest neighbor method to the discriminant analysis of the formalized data.

Index terms: big data, dichotomous data, discriminant analysis, nearest neighbor, variable selection.

I. INTRODUCTION

It is necessary to use the dichotomous data in the analysis of social phenomena and biological and medical diagnosis. One way to analyze such data is discrete discriminant analysis. The related studies are 'polynomial classification' based on the Linear Discriminant Function(LDF) and the 'discriminant difference criterion' proposed by [6]. As interest in discriminant analysis of big data has increased in recent years, research on the field is active. The use of Big Data is also essential for politics and social issues that are sensitive to public opinion. Classifying precisely which respondents belong to which group is very helpful for policy formulation. [16] suggested the intelligent Voice of customer (VOC) analyzing system, using opinion mining in Big Data. It is because the VOC data is made of very emotional contents by voice or text of informal style, and the volume of the VOC data are so big. These unstructured big data are difficult to store and analyze for human usage. Therefore, the organizations need automatic collecting, storing, classifying and analyzing system for unstructured big VOC data. [16]propose an intelligent VOC analyzing system based on opinion mining to discriminate the unstructured VOC data automatically and determine the polarity as well as the type of VOC. The data used in the discriminant analysis are either formal or informal, and discriminant analysis of informal data has been studied recently by various scholars (e.g., [12],[14],[5],[10]

Revised Manuscript Received on January 31, 2019.
Kyubark Shim, Department of Bigdata and Applied Statistics, Dongguk University, College of Science and Technology, Gyeongju, Korea.
* Corresponding author

II. POLYNOMIAL CLASSIFICATION AND VARIABLE SELECTION

Let $\mathbf{x}_j$, $j=1,2,\cdots,p$ be a dichotomous variable with 0 or 1. When the p-dimensional polynomial distribution is composed of 2^p reaction vectors $\mathbf{x}$ in each population M_i ($i=1,2$), the unbiased estimates of function $f_i(\mathbf{x})$ are as follows.

$f_i(\mathbf{x}) = n_i(\mathbf{x}) / n_i$, $i=1,2,$

where $n_i(\mathbf{x})$ is the number of $\mathbf{x}$ observed among the number of samples n_i extracted from population M_i, and $f_i(\mathbf{x})$ is the conditional discrete probability density function in population M_i.

If the estimated prior probability in population M_i is $\delta_1 = n_1 / n$ and $\delta_2 = n_2 / n$, respectively, where $n = n_1 + n_2$, then the estimated discriminant scores $g_i(\mathbf{x})$ are as follows.

$$g_i(\mathbf{x}) = \delta_i f_i(\mathbf{x}),\ i=1,2. \tag{2-1}$$

Assuming independent sample of size n_i from M_i ($i=1,2$), the sample based classification rule is given by

classify $\mathbf{x}$ to M_1 if $\hat{g}_1(\mathbf{x}) > g_2(\mathbf{x})$,

classify $\mathbf{x}$ to M_2 if $\hat{g}_1(\mathbf{x}) < g_2(\mathbf{x})$,

randomly allocate if $\hat{g}_1(\mathbf{x}) < g_2(\mathbf{x})$. (2-2)

Let G be the ordered-partition $G = \langle G_1, G_2 \rangle$ of sample space S, and D be assigned to M_i ($i=1,2$) under the necessary and sufficient condition to satisfy $X = \mathbf{x} \in G_i$, $i=1,2$. When the conditional probability of the misclassification is $t(G \mid \mathbf{x})$ under the condition that $X = \mathbf{x} \in G_i$, $i=1,2$ satisfies $t(G \mid \mathbf{x})$, $t(G \mid \mathbf{x})$ is as follows.

$$t(G \mid \mathbf{x}) = g_j(\mathbf{x}) / g(\mathbf{x}),\ i \neq j. \tag{2-3}$$

Averaging out over $\mathbf{x}$ yields the conditional error rate

$$t(G) = E\{t(G \mid \mathbf{x})\} = \sum_{G_1} g_2(\mathbf{x}) + \sum_{G_2} g_1(\mathbf{x}). \tag{2-4}$$

A rule is optimal if it minimizes the unconditional probability of misclassification.

There is a method of performing discriminant analysis using the difference in distance between the discriminant points. [6] proposed $(1-d^2)^n$ by computing the upper bound bias of each point in the sample of size n, where

$$d = \inf_{\mathbf{x}} \left| \sqrt{g_1(\mathbf{x}(i))} - \sqrt{g_2(\mathbf{x}(i))} \right| .$$

Here, d is the distance between discriminant points. A larger value of d is a good discrimination, while a smaller value results in a larger discrimination error. In this sense, d acts as a pseudo-distance between the discriminant scores. Given a classification rule under perfect stochastic conditions, d can be determined so that a smaller classification error is performed.

$$\max_{1\le j\le p} \max_{1\le i\le \binom{p}{j}} \min_{\mathbf{x}(i)} \left| \sqrt{g_1(\mathbf{x}(i))} - \sqrt{g_2(\mathbf{x}(i))} \right| , \qquad (2\text{-}5)$$

where $\mathbf{x}(i)$ represents a particular combination when j out of the available p variables are used, $j = 1, 2, \cdots, p$ and $i = 1, 2, \cdots \binom{p}{j}$. A sample based analogue (2-5) is

$$\max_{1\le j\le p} \max_{1\le h\le \binom{p}{j}} \min_{\mathbf{x}(h)} \left| \sqrt{N_1(\mathbf{x}(h))} - \sqrt{N_2(\mathbf{x}(h))} \right| , \qquad (2\text{-}6)$$

where $N_i(\mathbf{x}(h))$ is the number of sample observations from group M_i ($i = 1, 2$) having $\mathbf{X}(h) = \mathbf{x}(h)$.

The expression given by (2-6) needs to be amended to account for the dimensionality of the sample space induced by restricting consideration to $\mathbf{X}(h)$. When the number of used variables with multiple trends and correlations increased, the number of points in the sample space increased, and hence for a fixed total sample of size N the expected number of observations at any point $\mathbf{x}(h)$ decreases. We would expect that the difference $\left| \sqrt{N_1(\mathbf{x}(h))} - \sqrt{N_2(\mathbf{x}(h))} \right|$ will be small when both $\sqrt{N_1(\mathbf{x}(h))}$ and $\sqrt{N_2(\mathbf{x}(h))}$ are expected to be small. Scaling the frequency estimate thus seems appropriate, and indeed the authors consider the following expression

$$\max_{1\le j\le p} \max_{1\le h\le \binom{p}{j}} \min_{\mathbf{x}(h)} \left| \frac{\sqrt{N_1(\mathbf{x}(h))} - \sqrt{N_2(\mathbf{x}(h))}}{\left(M_j l_j\right)^{\frac{1}{2}}} \right| , \qquad (2\text{-}7)$$

where l_j is the number of levels of variable j and where the product is over those variables other than those in $\mathbf{X}(h)$. In particular, if $\mathbf{x}$ is a multivariate binary vector of dimension p, then $M_j l_j = 2^{p-k}$.

In applying (2-7) it is generally found that too many variables are selected. Because of this tendency a variation can be utilized, namely :

$$\max_{1\le j\le p} \max_{1\le h\le \binom{p}{j}} \operatorname{avg}_{\mathbf{x}(h)} \left| \frac{\sqrt{N_1(\mathbf{x}(h))} - \sqrt{N_2(\mathbf{x}(h))}}{\left(M_j l_j\right)^{\frac{1}{2}}} \right| . \qquad (2\text{-}8)$$

In words, (2-8) chooses that subset of variables that maximizes the scaled average value of $\sqrt{N_1(\mathbf{x}(h))} - \sqrt{N_2(\mathbf{x}(h))}$.

III. NEAREST NEIGHBOR OF ORDER R AND THE VARIABLE SELECTION

Nearest neighbor of order r can be used as a way to reduce the error of polynomial classification rule. This method is that when using a sample-based likelihood ratio procedure for classifying a particular response vector $\mathbf{x}$, all responses differing from $\mathbf{x}$ in no more than r components are incorporated into the rule. Let us denote the reaction vector of degree $r \le 1$ in all response vectors $\mathbf{x}$ as follows.

$$T_j = \left\{ \mathbf{y}_j \mid (\mathbf{x}-\mathbf{y}_j)(\mathbf{x}-\mathbf{y}_j)^T \le r \right\}, \; j = 1, 2, \cdots, p . \qquad (3\text{-}1)$$

Note that T_j is merely the set of responses $\{\mathbf{y}_j\}$ having the property that each of its elements differs in no more than r components from $\mathbf{x}$.

Assuming independent samples of size n_i from M_i ($i = 1, 2$), the nearest neighbor of order $r \le 1$ the rule is given by

classify $\mathbf{x}$ into M_1 if $\hat{\delta} \sum_{T_j} \frac{n_1(\mathbf{y}_j)}{n_1} > (1-\hat{\delta}) \sum_{T_j} \frac{n_2(\mathbf{y}_j)}{n_2}$,

classify $\mathbf{x}$ into M_2 if $\hat{\delta} \sum_{T_j} \frac{n_1(\mathbf{y}_j)}{n_1} < (1-\hat{\delta}) \sum_{T_j} \frac{n_2(\mathbf{y}_j)}{n_2}$,

and randomly allocate if $\hat{\delta} \sum_{T_j} \frac{n_1(\mathbf{y}_j)}{n_1} = (1-\hat{\delta}) \sum_{T_j} \frac{n_2(\mathbf{y}_j)}{n_2}$

(3-2)

Note that in the nearest neighbor rule, we consider only the rule determined by nearest neighbor of order $r = 1$. To make the nearest neighbor rule operational, it is required that the set of near neighbors for each set to be specified. For example, if $\mathbf{x} = (1 \quad 1 \quad 1 \quad 1)$ is given, the near neighbors of order $r = 1$ of $\mathbf{x}$ are given by

$T_{1111} = \{(0111), (1011), (1101), (1110)\}$.

In multinomial classification rule, only the reaction vector itself can be analyzed, but it is an advantage of this discrimination method, that it is possible to analyze all the responses within a predetermined order range. In the analysis using nearest neighbor rule, the response frequency is increased by multiplying the frequency of each group by the number of components contained in $\mathbf{y}_j$. If the number of components is m, the frequency of each group is classified using nearest neighbor rule is $mN_1(\mathbf{x}(h))$ and $mN_2(\mathbf{x}(h))$, respectively. Therefore, in the case of nearest neighbor rule of order $r = 1$, the following new equation is used for the variable selection method.

$$\max_{1\le j\le p}\max_{1\le h\le\binom{p}{j}}\min_{\mathbf{x}(h)}\left|\frac{\sqrt{mN_1(\mathbf{x}(h))}-\sqrt{mN_2(\mathbf{x}(h))}}{\left(M_j l_j\right)^{\frac{1}{2}}}\right|, \quad \text{(3-3)}$$

$$\max_{1\le j\le p}\max_{1\le h\le\binom{p}{j}}\operatorname{avg}_{\mathbf{x}(h)}\left|\frac{\sqrt{mN_1(\mathbf{x}(h))}-\sqrt{mN_2(\mathbf{x}(h))}}{\left(M_j l_j\right)^{\frac{1}{2}}}\right|. \quad \text{(3-4)}$$

Equation (3-3) will calculate the distance between one point with the shortest inter-group distance between group 1 and group 2. Therefore, when equation (3-3) is used there is a possibility that too many variables are selected. Equation (3-4) is an effective method because it calculates the distance between the mean values of the two groups.

IV. CONCLUSION

In this paper, we propose a variable selection method using the nearest neighbor rule of order $r=1$ in 2-group discrete discriminant analysis. There is a method to select the discriminant variable by using the scaled minimum value (3-3) and the scaled average value (3-4). The accuracy of these two methods varies depending on the situations, and can be judged by calculating and comparing the error rate of each misclassification.

ACKNOWLEDGMENT

This work was supported by the Special Fund for Overseas Training of Dongguk University in 2017.

REFERENCES

1. Ali, A., & Haseeb, M. (2019). Radio frequency identification (RFID) technology as a strategic tool towards higher performance of supply chain operations in textile and apparel industry of Malaysia. Uncertain Supply Chain Management, 7(2), 215-226.
2. Awang, Z., Ahmed, U., Hoque, A. S. M. M., Siddiqui, B. A., Dahri, A. S., and Muda, H. (2017). The Mediating Role of Meaningful Work in the Relationship Between Career Growth Opportunities and Work Engagement, International Academic Confernce on Business and Economics (IACBE 2017), Faculty of Economics and Management Sciences (FESP), Universiti Sultan Zainal Abidin (UniSZA), October 07-08
3. Celeux G., Mkhadri A.(1992): Discrete regularized discriminant analysis. Statistics and Computing 2(3): 143-151.
4. Esfahani, H., Tavasoli, K & Jabbarzadeh, A. (2019). Big data and social media: A scientometrics analysis.International Journal of Data and Network Science, 3(3): 145-164.
5. Ferreira A.S.(2010), A Comparative Study on Discrete Discriminant Analysis through a Hierarchical Coupling Approach. In Classification as a Tool for Research, Studies in Classification, Data Analysis, and Knowledge Organization; Hermann Locarek-Junge, Claus Weihs (Eds.), Springer-Verlag, Heidelberg-Berlin: 137~145.
6. Glick, N.(1973), "Sample-Based Multinomial Classification", Biometrics, 29, 241~256.
7. Haseeb, M., Abidin, I. S. Z., Hye, Q. M. A., & Hartani, N. H. (2018). The Impact of Renewable Energy on Economic Well-Being of Malaysia: Fresh Evidence from Auto Regressive Distributed Lag Bound Testing Approach. International Journal of Energy Economics and Policy, 9(1), 269-275.
8. Haseeb., H. Z., G. Hartani., N.H., Pahi., M.H. Nadeem., H. . (2019). Environmental Analysis of the Effect of Population Growth Rate on Supply Chain Performance and Economic Growth of Indonesia. Ekoloji, 28(107).
9. Suryanto, T., Haseeb, M., & Hartani, N. H. (2018). The Correlates of Developing Green Supply Chain Management Practices: Firms Level Analysis in Malaysia. International Journal of Supply Chain Management, 7(5), 316.
10. Goldstein, M. and Rabinowitz, M.(1975), "Selection of Variate for the Two-Group Multinomial Classification Problem", Journal of American Statistics Association, 70, 776~781.
11. Goldstein, M. and Dillon, W.R.(1978), Discrete Discriminant Analysis, John Wiley and Sons, New York.
12. Goldstein, M. and Dillon, W.R.(1984), Multivariate Analysis Methods and Application, John Wiley and Sons, New York.
13. Karthick, K., Premkumar, M., Manikandan, R., & Cristin, R. (2018). Survey of Image Processing Based Applications in AMR. Review of Computer Engineering Research, 5(1), 12-19.
14. Hills, M.(1967), "Discrimination and Allocation with Discrete Data", Applied Statistics, 16,237~250.
15. İnel, M. (2019). An empirical study on measurement of efficiency of digital transformation by using data envelopment analysis. Management Science Letters, 9(4), 549-556.
16. Kim, Y.S. and Jeong, S.Y.(2013), "Intelligent VOC Analyzing System Using Opinion Mining", Journal of Intelligence and Information Systems, 3, 113~125.
17. Marques, A, Ferreira, A.S. and Margarida, G.M.S.(2013) "Selection of variables in Discrete Discriminant Analysis", Biometrical Letters, 50, 1~14.
18. Perez-de-la-Cruz G, Eslava-Gomez G (2016) Discriminant analysis with Gaussian graphical tree models. AStA Adv Stat Anal 100: 161-187.

Kyubark Shim, received the M.S. and the Ph.D. degrees in Statistics from the Dongguk University in 1986 and 1993, respectively. He is a Professor in the Department of Bigdata and Applied Statistics at Dongguk University at Gyeongju, Korea. His current research interests include Computational Statistics, Reliability

19.