Two-step Variable Screening Method for the Mahalanobis-Taguchi Method with Small Training Data

YUTO UEMURA* Non-member,
ARATA SUZUKI*[†] Non-member

DAISUKE FUJITA* Non-member

(Received October 12, 2017, revised March 18, 2018)

Abstract: This paper proposes a method that applies the Mahalanobis-Taguchi (MT) method in situations where the number of variables exceeds the number of training samples. In the proposed method, the diagnostic effectiveness of the variables is first evaluated using signal-to-noise (SN) ratio, and the MT method is then applied using only those variables with high diagnostic effectiveness. Accordingly, the variables are selected via a two-step process: (1) variables are selected before creating the unit space and (2) conventional variable selection is carried out using the MT method. The results of experimental evaluation of the effectiveness of the proposed method using four different datasets from the UCI Machine Learning Repository indicate that its diagnostic performance is superior to that of the RT method.

Keywords: multivariate analysis, Mahalanobis-Taguchi system, signal-to-noise ratio, Mahalanobis distance, variable selection, small training data

1. Introduction

Multivariate diagnostic techniques are increasingly being used in the medical and industrial fields. Among the multivariate diagnostic techniques, the Mahalanobis-Taguchi System (MTS) has been gaining attention because of its highly practical applicability [1], [2]. MTS is the generic name for a group of methods that includes the Mahalanobis-Taguchi (MT), T, and RT methods. Among these methods, the MT method combines the Mahalanobis distance with the Taguchi SN ratio, and has been applied in many diagnostic case studies [3], [4], [5], [6], [7]. In the MT method, a Mahalanobis space called the unit space is created from the normal (homogeneous) group and diagnosis is performed with this unit space as the reference. Unlike conventional classification methods, it does not take into account the class distribution, but rather performs the diagnosis based on matching to one class (unit space) only. Because of these characteristics, its effectiveness in practical applications is considered high.

The effectiveness of the MT method is evident from the large number of case studies available, but there are problems associated with it. One such problem is multicollinearity, which occurs when the number of training samples is less than the number of variables, or when there are linear associations between two or more variables. The former situation is called perfect multicollinearity; the MT method cannot be applied in such a situation because the Mahalanobis space cannot be created. However, in recent times,

building effective diagnostic models using a limited amount of sample data from the initial production phase has become crucial because of short product development cycles.

Thus, this paper proposes a method that applies the MT method in situations where the number of training samples is less than the number of variables. In the proposed method, application of the MT method is made possible by selecting useful diagnostic variables that are fewer in number than the training samples. Accordingly, selection of variables in the proposed method comprises two steps: (1) variable selection before creating the unit space and (2) conventional variable selection using the MT method. Further, this paper proposes using the SN ratio for diagnostic purposes (diagnostic SNR) in selecting (screening) variables before creating the unit space.

The proposed method enables the MT method to be applied in situations where perfect multicollinearity exists by using already proposed simple screening approach [8]. Engineers familiar with the MT method can easily apply the proposed method because it is similar to the conventional MT method except for the screening. Use of the MT method in the field is difficult because the common optimizing techniques used in variable selection in the MT method call for complex calculations [9], [10]. Moreover, its variable selection method is not applicable in perfect multicollinearity situations. In MTS, two methods usually adopted in situations of perfect multicollinearity are the T method and the RT method [11], [12], [13], [14]. The T method is used for targets that are near the center of the unit space, and is not suitable for two-class (good/negative) diagnosis. The RT method is used when all the variables are in the same unit of measure, and nondimensionalization is necessary for vari-

^{§§} Corresponding: aratas@sys.wakayama-u.ac.jp

* Faculty of Systems Engineering, Wakayama University
930 Sakaedani, Wakayama city, 640-8510, Japan
aratas@sys.wakayama-u.ac.jp

ables in different units. Although dimension reduction can be considered for handling perfect multicollinearity, identification of useful variables for diagnosis is not possible in such an approach. A significant benefit of the proposed method is that it enables identification of useful variables using SN ratio in a manner similar to the conventional MT method.

2. The MT and RT Methods

2.1 The MT Method In the MT method, a Mahalanobis space, called a unit space, is created from the normal (homogeneous) group, and diagnosis is performed based on this unit space. The normal group refers to a group having common characteristics, such as a collection of samples that have passed the quality inspection tests, or a group of individuals who have been diagnosed as healthy in health checkups.

The procedure applied to create a diagnostic model using the MT method is described below. First, training samples are collected, from which the normal group is then selected. The unit space is then calculated from this normal group. Finally, variable selection is performed based on the samples that do not belong to the unit space-collectively called the signal space. The procedure followed in the MT method is as follows.

2.2 Procedure followed in the MT method Select n training samples with k variables from the normal group. Assuming the training sample to be x_{ij} ($i = 1, 2, \dots, n; j = 1, 2, \dots, k$), standardize it using the following equation:

$$X_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j} \quad (1)$$

where $\bar{x}_j$ and σ_j are the mean and standard deviation of the j th variable, respectively. Using these values, calculate correlation matrix $\mathbf{R}$ and its inverse $\mathbf{A}$. The $\bar{x}_j$, σ_j and $\mathbf{A}$ thus obtained for all the variables constitute the unit space. Assuming the diagnosis data standardized using equation (1) to be $\mathbf{Y} = (Y_1, Y_2, \dots, Y_k)$, the Mahalanobis distance (MD) can be derived from the following equation:

$$MD = \frac{1}{k} \mathbf{Y} \mathbf{A} \mathbf{Y}^T \quad (2)$$

where MD is equal to zero if the diagnosis data coincide with the mean $\bar{x}_j$ of the unit space. Using this property, the diagnosis is performed based on the MD value.

Next, the variables useful in diagnosis are selected from the unit space for all the variables. A two-level orthogonal array is used in the variable selection process. In a two-level orthogonal array, each factor has two levels and the factors constitute the columns in the matrix (that is, the rows of the array represent the level combinations, and columns the factors). Table 1 shows the L_8 orthogonal array as an example of a two-level orthogonal array. In this table, a combination of up to seven variables is possible. The total number of possible combinations is $2^7 = 128$, but in an orthogonal array, this is reduced to eight combinations.

The level determines whether the variable will be selected or not. Assume that the variables marked level 1

Table 1: L_8 Orthogonal array.

	Variables						
	#1	#2	#3	#4	#5	#6	#7
Test 1	1	1	1	1	1	1	1
Test 2	1	1	1	2	2	2	2
Test 3	1	2	2	1	1	2	2
Test 4	1	2	2	2	2	1	1
Test 5	2	1	2	1	2	1	2
Test 6	2	1	2	2	1	2	1
Test 7	2	2	1	1	2	2	1
Test 8	2	2	1	2	1	1	2

are selected, and those marked level 2 are not. Using Table 1 as an example, in Test 1, all the variables are used, but in Test 2 only variables 1, 2, and 3 are used. The unit space is created according to these combinations, and the Mahalanobis distance of the training sample for the signal space is calculated. Because the diagnosis accuracy is better with larger Mahalanobis distance for the training sample of the signal space, the test results are evaluated using the larger-the-better SN ratio property given by the following equation:

$$\eta_{LTB}^l = -10 \log_{10} \frac{1}{m} \sum_{l=1}^m \frac{1}{MD_l^2} \quad (3)$$

where MD_l ($l = 1, 2, \dots, m$) is the Mahalanobis distance of the training sample of the signal space. The SN ratio for each variable is calculated based on the results of the tests. If the mean SN ratio η_1 obtained from level 1 tests (where variables are used) and the mean SN ratio η_2 obtained from level 2 tests (where variables are not used) satisfy the condition $\eta_1 < \eta_2$, then the variable can be considered as not useful. The unit space optimized based on such selection of variables useful for diagnosis, forms the basis of the diagnostic model. It is to be noted that η_1 is the mean of the SN ratios for variable 1 obtained from Tests 1 through 4.

2.3 Short summary of the RT method As stated above, one of the problems affecting the MT method is perfect multicollinearity. The method proposed in this paper addresses the situation in which the number of unit space samples is less than the number of variables. Measures to handle such a situation have been proposed in MTS. One such measure, the RT method, is similar to the MT method. In the RT method, a unit space is calculated and the diagnosis is based on the distance from this unit space. The most significant difference between the RT and MT methods is that in the RT method, all the variables are reduced to just two variables, the standard SN ratio and the sensitivity. This eliminates the effect of perfect multicollinearity. A detailed procedure for the RT method can be found in reference [12].

3. Two Stage Variable Selection Method

This section describes the MT method, in which the two-step variable selection process comprising screening based on the diagnostic SNR and selection based on the orthogonal array, was adopted. In the proposed MT method, first,

Figure 1: Flow of the proposed method.

screening of the variables is performed using the diagnostic SNR, and then the conventional MT method is applied to the screened variables. The flowchart of the MT method to which the two-step variable selection process has been applied is shown in Fig. 1.

3.1 Diagnostic SNR The MT method performs diagnosis using the unit space created from the normal group. Because the diagnosis is based on the unit space, its uniformity is very important. Moreover, in order to improve the accuracy of the diagnosis, a unit space with a large Mahalanobis distance for the signal space is desirable.

In summary, the following are deemed essential for high diagnosis accuracy: (1) small variability for samples in the unit space, and (2) large distance for the samples in the signal space. The diagnostic SNR $\eta_{sc}^{(j)}$ of j th variable given by the following equation was developed to evaluate these properties:

$$\eta_{sc}^{(j)} = 10 \log_{10} \frac{\frac{1}{m} \sum_{l=1}^m (\bar{x}_j - y_{lj})^2}{\sigma_j^2} \quad (4)$$

where $\bar{x}_j$ and σ_j^2 are the mean and variance of the j th variable of the training samples in the unit space, respectively, and y_{lj} and m ($l = 1, 2, \dots, m$) are the variables of the training sample and the number of samples in the signal space, respectively.

3.2 Screening step Screening of the variables is conducted based on the diagnostic SNR, which is calculated for all the variables. The variables are selected according to the descending order of the diagnostic SNR such that the number of variables selected does not exceed the number of training samples. It is essential that the number of variables selected be at least two less than the number of training samples. Thus, in this paper, in accordance with the case studies of the MT method, the number was set to approximately one-third the number of training samples [15].

The Mahalanobis distance is a distance measure that is based on the correlation between two variables. Although

the proposed SN ratio does not take into account the correlation between two variables, emphasizing the importance of uniformity of the unit space used in the MT method, and considering that it is a two-class (good/negative) diagnosis, the authors believe that variables are selected in conformance with the concept in the MT method. Moreover, as the conventional MT method is applied after screening of the variables, the unit space is created by considering the correlation between the screened variables, as in the conventional MT method.

4. Experiment

4.1 Experiment method Diagnostic experiments were conducted in which datasets from the UCI Machine Learning Repository were used to confirm the accuracy of the proposed method. The following four datasets with a relatively large number of variables were used:

- Wisconsin Diagnostic Breast Cancer (WDBC)
- Parkinsons
- Cleveland Heart Disease (CHD)
- Cardiotocography

In the experiments, the unit space was calculated from the healthy group assuming the group to be homogeneous. Although the Cardiotocography dataset had three classes—specifically, Normal, Suspect, and Pathologic—in the experiments for this study, only the data for the two classes Normal and Pathologic were used. An outline of the data used in the experiments is shown in Table 2.

Training samples were selected from the dataset after removing those variables that were not suitable (standard deviation = 0) for the MT method, such that the number of training samples was either the same as the number of variables or less. In the resulting datasets, the number of training samples in the unit space and the signal space was 120 (four groups with 30 samples per group) in the WDBC, CHD, and Cardiotocography datasets, and 40 (two groups with 20 samples per group) in the Parkinson dataset. These groups were divided randomly. The number of variables selected in the screening step was set to approximately one-third the number of samples. The number of training samples and the number of variables used are shown in Table 3. Regarding the orthogonal array used in the experiments, a two-level L_{12} orthogonal array was used for the WDBC, CHD, and Cardiotocography datasets, while a two-level L_8 orthogonal array was used for the Parkinsons dataset, in accordance with the number of variables for the respective datasets.

The evaluation of diagnostic performance was based on the true positive rate and the false positive rate and was conducted through cross-validation using the divided groups. Assuming the unit space data to be normal and the signal space data to be abnormal, the true positive rate was calculated as the proportion of abnormal data correctly diagnosed as abnormal, whereas the false positive rate was calculated as the proportion of normal data incorrectly diagnosed as

Figure 2: Diagnosis accuracy of the proposed method and the RT method.

Table 2: Summary of datasets.

	Unit Space	Signal Space
WDBC	Benign	Malignant
Parkinsons	Healthy	Parkinson
CHD	Normal	Abnormal
Cardiotocography	Normal	Disease

Table 3: Sample data amount and variable.

	Unit Space	Signal Space	Variable
WDBC	30×4	30×4	10 (/30)
Parkinsons	20×2	20×2	7 (/22)
CHD	30×4	30×4	10 (/26)
Cardiotocography	30×4	30×4	10 (/22)

abnormal. The closer the true positive rate to 100% and the closer the false negative rate to 0%, the better is the diagnostic performance.

The RT method, which is not affected by perfect multicollinearity, was used for comparison. The nondimensionalization of the variables was performed using standard deviation. Moreover, in order to correctly compare the performance of the two methods, a threshold was set such that the false positive ratio was the same in both the proposed method and the RT method. The performance was evaluated based on the true positive rate achieved under these conditions.

4.2 Experimental results Figure 2 shows the mean true positive ratio and the mean false positive ratio for the different groups in each of the datasets. The true positive ratio for WDBC was 96.7% in the proposed method and 86.9% in the RT method, for Parkinsons it was 87.5% in the proposed method and 58.2% in the RT method, for CHD it was 57.2% in the proposed method and 19.8% in the RT method, and for Cardiotocography it was 99.2% in the proposed method and 76.3% in the RT method. The false positive ratio for WDBC was 25.3% in the proposed method and 7.5% in the RT method, for Parkinsons it was 25.0% in the proposed method and 7.5% in the RT method, for CHD it was 21.9% in the proposed method and 8.1% in the RT method, and for Cardiotocography it was 25.0% in the proposed method and 8.1% in the RT method. The threshold was set at a 1% significance level for the chi-squared distribution.

Figure 3: Diagnosis accuracy of both methods with adjusting threshold.

Figure 3 shows the mean true positive ratio when the false positive ratio of the proposed method was equal to that of the RT method. The diagnosis accuracy for WDBC was 96.7% in the proposed method and 95.2% in the RT method, for Parkinsons it was 87.5% in the proposed method and 74.1% in the RT method, for CHD it was 57.2% in the proposed method and 39.8% in the RT method, and for Cardiotocography it was 99.2% in the proposed method and 94% in the RT method.

4.3 Discussion The effect of the proposed two-step variable selection process was verified experimentally using the WDBC, Parkinsons, CHD, and Cardiotocography datasets. The results obtained in the experiments demonstrate that for all the datasets the diagnosis accuracy of the proposed method is better than that of the RT method, when the false positive rates of the methods are the same. The authors believe that the proposed method has better diagnosis accuracy than the RT method, and it can likely be an effective MT method that can be applied to situations with perfect multicollinearity.

One issue that remains to be studied is the setting of thresholds, which was not discussed in this paper. The threshold for the experiments in Fig. 2 was set at 1% significance level for the chi-squared distribution. The threshold for the experiments in Fig. 3 was set such that the false positive rates of the two methods are equal. The results suggest that the true positive rate of the proposed method does not decrease even when the threshold is set lower than the false positive rate. The threshold for the experiments in Fig. 2 was set using the unit space only; however, the authors believe that a threshold considering the signal space needs to

be set. Thus, development of an appropriate threshold selection method is an important topic that needs to be studied in the future.

5. Conclusions and Future Work

This paper proposed a method that applies the MT method after screening the variables as a measure to handle perfect multicollinearity in situations where the number of variables exceeds the number of training samples. The proposed method employs a two-step variable selection process with a newly developed variable called diagnostic SNR that is used in combination with conventional orthogonal array-based variable selection. The results of evaluations conducted of the diagnosis accuracy of the proposed method using four different datasets indicate that its performance is superior to that of the RT method.

In recent times, the product development cycle has been getting increasingly shorter. Thus, the authors believe that building accurate diagnostic models based on limited sample data from the initial development phase, which the proposed method is highly capable of doing, is extremely important. Methods for setting appropriate thresholds and for screening based on correlation between variables will be investigated in future work.

References

- [1] G. Taguchi and R. Jugulum, "The Mahalanobis-Taguchi strategy: A pattern technology system", Wiley, New York, USA, 2002.
- [2] W. H. Woodall, R. Koudelik, K-L. Tsui, S. B. Kim, Z. G. Stoumbos, and C. P. Carvounis, "A review and analysis of the Mahalanobis-Taguchi system", *Technometrics*, Vol.45, No.1, pp.1-15, 2003. DOI: 10.1198/004017002188618626
- [3] H. Yuki and D. Kai, "Application of Mahalanobis-Taguchi method to identification of the initial motion part in acoustic emission waveforms", *Transactions of the Japan Society of Mechanical Engineers Series C*, Vol.79, No.797, pp.145-149, 2013. (in Japanese) DOI: 10.1299/kikaic.79.145
- [4] C-W. Kim, K-C. CHANG, and S. Kitauchi, "Anomaly detection from changes in statistical patterns of structural dynamic characteristics of a simply supported steel-truss bridge", *Journal of Japan Society of Civil Engineers, Ser. A2*, Vol. 70, No. 2 pp. I.63-I.72, 2014. (in Japanese) DOI: 10.2208/jscejam.70.I.63
- [5] M. Mita, A. Homma, Y. Mita, T. Nishinaka, H. Sumikura, K. Ohnuma, E. Tatsumi, Y. Taenaka, A. Funakubo, and Y. Fukui, "A Study of anatomical compatibility of VAD using Mahalanobis distance", *Journal of Life Support Engineering*, Vol.26, No.2, pp.57-63, 2014. (in Japanese) DOI: 0.5136/lifesupport.26.57
- [6] C. Li, J. Yuan and P. Gao, "Risk decision-making based on Mahalanobis-Taguchi system and grey cumulative prospect theory for enterprise information investment", *Intelligent Decision Technologies*, Vol.10, No.1, pp.49-58, 2016. DOI: 10.3233/IDT-150236
- [7] B. Buenviaje, J. E. Bischoff, R. A. Roncace, and C. J. Willy, "Mahalanobis-Taguchi system to identify preindicators of delirium in the ICU", *IEEE Journal of Biomedical and Health Informatics*, Vol.20, No.4, pp.1205-1212, 2016. DOI: 10.1109/JBHI.2015.2434949
- [8] D. Fujita, Y. Uemura, and A. Suzuki, "A Simple Variable Screening Method for the Mahalanobis Taguchi Method", *Journal of the Institute of Industrial Applications Engineers*, Vol.5, No.3, pp.111-117, 2017. DOI: 10.12792/JIAE.5.111
- [9] E. Reséndiz and C. A. Rull-Flores, "Mahalanobis-Taguchi system applied to variable selection in automotive pedals components using Gompertz binary particle swarm optimization", *Expert Systems with Applications*, Vol.40, No.7, pp.2361-2365, 2013. DOI: 10.1016/j.eswa.2012.10.049
- [10] A. S. Iquebal, A. Pal, D. Ceglarek, and M. K. Tiwari, "Enhancement of Mahalanobis-Taguchi system via rough sets based feature selection", *Expert Systems with Applications*, Vol.41, No.17, pp.8003-8015, 2014. DOI: 10.1016/j.eswa.2014.06.019
- [11] E. A. Cudney, P. A. Shah, and R. Kestle, "Predicting vehicle cost using the T-method", *Int. J. of Product Development*, Vol.12, No.3, pp.311-323, 2010. DOI: 10.1504/IJPD.2010.036393
- [12] F. Huda, I. Kajiwar, N. Hosoya, and S. Kawamura, "Bolt loosening analysis and diagnosis by non-contact laser excitation vibration tests", *Mechanical Systems and Signal Processing*, Vol.40, No.2, pp.589-604, 2013. DOI: 10.1016/j.ymssp.2013.05.023
- [13] Arata Suzuki, "Inverse-model-based cuffless blood pressure estimation using a single photoplethysmography sensor", *Journal of Engineering in Medicine*, Vol.229, No.7, pp.499-505, 2015. DOI: 10.1177/0954411915587957
- [14] S. Negishi, Y. Morimoto, S. Takayama, and A. Ishigame, "Daily Peak Load Forecasting by Taguchi's T Method", *IEEE Transactions on Electronics, Information and Systems*, Vol.136, No.6, pp.794-801, 2016. (in Japanese) DOI: 10.1541/ieejieiss.136.794
- [15] H. Aman, N. Mochiduki and H. Yamada, "A Model for Detecting Cost-Prone Classes Based on Mahalanobis-Taguchi Method", *IEICE transactions on information and systems*, Vol.89, No.4, pp.1347-1358, 2006.

Yuto Uemura (Non-member) received the B.E. degree in Systems Engineering from Wakayama University, Japan, in 2016. He is working toward the M.E. degree at Wakayama University. His research interests include statistical data analysis, and biomedical data analysis.

Daisuke Fujita (Non-member) received the M.E. degree in Systems Engineering from Wakayama University, Japan, in 2014. He is currently pursuing a Ph.D. degree in systems engineering at Wakayama University. His research interests include machine learning, and biomedical data analysis. Mr. Fujita is a member of JSMBE.

Arata Suzuki (Non-member) received the Ph.D. degree in information science from Nara Institute of Science and Technology (NAIST), Japan, in 2006. He is currently a lecturer of Wakayama University. His research interests include statistical data analysis, Taguchi robust engineering and biomedical informatics. Dr. Suzuki is a member of JSMBE, ISCIE, JSQC

and IEEEJ.