Application of Mahalanobis Taguchi System for Analysis of Multivariate System

(Mahalanobis Taguchi System을 이용한 다변량 시스템의 해석에 관한 연구)

홍정의* 김용범**

Abstract

Mahalanobis Taguchi System (MTS) is developed by Genishi Taguchi as a part of his quality engineering methodology. The basic idea of Taguchi's quality engineering is looking for the way of effectiveness of analyzing multivariate system. In the MTS, with the standardized variables of healthy normal data, Mahalanobis Distance(MD) calculated and that can be discriminate between normal and abnormal objects. If this discrimination process is successful, next step is optimization which is try to reduce number of attributes by neglecting less effective attributes to MD. Orthogonal Array (OA) and Signal to Noise ratio (S/N) are used to evaluate the amount contribution of each attribute to the MD. Wisconsin Breast Cancer study, from machining learning repository at University of California at Irvine, used for examining the discriminant ability of MTS.

Keyword : Mahalanobis Taguchi System, Pattern Recognition

1.Introduction

Mahalanobis Distance (MD) was introduced as a statistical tool by Professor P. C. Mahalanobis in 1930 to distinguish patterns of a certain group from another group. Genichi Taguchi later introduced the Mahalanobis-Taguchi System (MTS) to define the reference group and measure the individual subsets[1]. MTS is a very economic approach for multidimensional pattern recognition systems.

* 충주대학교 산업경영공학과

** 충주대학교 경영학과

A pattern is defined as opposite of a chaos. For example, a pattern could be a fingerprint image, a handwritten cursive word, a human face, or a speech signal. A pattern can be distinguished as supervised and unsupervised. For the supervised classification, the input pattern has a predefined class. Pattern recognition is the study of how to observe and distinguish patterns of interest, and make reasonable decisions about the categories of pattern [2].

In multidimensional systems, it is necessary to reduce the number of variables by neglecting the variables that have little or no affect on the measurement function. There are numerous different approaches conducted previously such as linear discriminant analysis, logistic regression decision trees, and neural networks. [3]

The goal of this study is to looking for the discrimination ability of MTS. Wisconsin breast cancer data, from Machining learning repository at university of California at Irvine, which has nine attributes and one class used for this study. Section 2 descirbes brief literature review. Section 3 describes details of MTS. Section 4 shows simulation results. Section 5 summarizes the conclusion.

2.Literature Review

Considerable research is available utilizing Mahalanobis Distance to determine similarities of values from known and unknown samples. Existing research also uses the Mahalanobis Taguchi System for prediction and diagnosis which illustrates the methodology's accuracy and effectiveness. However, little is presented to compare the accuracy and effectiveness of the Mahalanobis Taguchi System versus other methodologies.

Taguchi utilized the Mahalanobis Taguchi System for diagnosis and pattern recognition. His research discussed a case study involving liver disease diagnosis in Tokyo, Japan using fifteen variables. In his research, he developed an eight-step procedure titled "Mahalanobis Distance for Diagnosis and Pattern Recognition System Optimization Procedure". [4]

Lande also conducted research using Mahalanobis Distance to evaluate potential habitats for large carnivores in Scandinavia. The species involved included bears, wolves, lynx, and wolverines. The variables used included landcover, human density, infrastructure, and prey density. The results of the study were used to determine which areas were suitable for each species. This research considered a different field for application with respect to habitats and the environment. [5]

Mahalanobis Distance was also used to maximize productivity in a new manufacturing control system by Hayashi et al.. The research used Mahalanobis Distance as a core to their manufacturing control system because of the methods ability to recognize patterns. The new system detected deviations from normal productivity much earlier and enabled root cause identification and prioritized resolution. [6]

Asada used the Mahalanobis-Taguchi System to forecast the yield of wafers in semiconductor manufacturing. Yield of wafers is determined by the variability of electrical characteristics and dust. The research focused on one wafer product that had a high yield. Mahalanobis Distances were calculated on various wafers to compare the relationship between yield and distance. The signal-to-noise ratios were used to indicate the capability of forecasting and the effect of the parameters. This research showed the applicability of the Mahalanobis Distance to forecasting defective components. [7]

Pattern recognition using Mahalanobis distance was demonstrated in the work of Wu. In this research, pattern recognition was used to diagnose human health. The results of tests from a regular physical check-up were used as the characteristics. The correlation between different tests was shown. Mahalanobis Distance was used to summarize the multi-dimensional characteristics into one scale. In this research the base point was difficult to define since it was a healthy person. People who were judged to be healthy for the past two years were considered to be healthy. The research considered diagnosis of liver function with the objective to forecast serious disease until the next check-up. The approach provided a more efficient method that also avoided inhuman treatment that previously used double blind tests. [8]

The research in this paper compares the accuracy and effectiveness of the Mahalanobis Taguchi System This research will provide a method for understanding and, therefore, meeting consumer requirements. Consequently, this will result in higher consumer satisfaction and lower costs by providing only characteristics that are key to the consumer.

3.MahalanobisTaguchiSystem

The Mahalanobis Taguchi System (MTS) is a pattern recognition technology that aids in quantitative decisions by constructing a multivariate measurement scale using a data analytic method. The main objective of MTS is to make accurate predictions in multidimensional systems by constructing a measurement scale [2]. The patterns of observations in multidimensional system highly depend on the correlation structure of variable in the system. One can make the wrong decision about the patterns if each variable is looked at separately without considering the correlation structure. To construct a multidimensional measurement scale, it is important to have a distance measure. The distance measure is based on the correlation between the variable and the different patterns that could be identified and analyzed respect to base or reference point.

In the MTS, the Mahalanobis space is calculated using the standardized variables of healthy or normal data. The Mahalanobis space can be used to discriminate between normal and abnormal objects. Once this MS is established, the number of attributes used is reduced using Orthogonal Array (OA) and Signal to Noise ratio (SN) to evaluate the contribution of each attribute. Each row of the OA determines a subset of the original system by the including and excluding that attribute of system.

To apply MTS, the first step is to define and sample normal observations to construct a reference space, which is called Mahalanobis Space (MS). Next, the Mahalanobis Distance (MD) is calculated. MD has the ability to differentiate a normal group from an abnormal group. MD summarizes multidimensional characteristics into a one dimensional scale by calculating the relationship between each characteristic.

If the MD can not effectively differentiate a normal and an abnormal group, then new samples should be taken and variables need to be established to build the MS. The next step is applying an orthogonal array and SN ratio to evaluate the contribution of each attribute and reduce the number of attributes. With the reduced number of attributes, the processing time and data collection cost can be reduced. The general procedures of MTS are described as follows [9]:

The first step in MTS is to construct a measurement scale using the MS as a reference. To construct a measurement scale, a data set of the normal observations needs to be collected. The collected normal observations are standardized using equation (1).

$$Z_i = \frac{X_i - m}{\sigma} \qquad (1)$$

where,

m, mean of the attribute,
σ, standard deviation of attribute,
Zi , standardized vector,
m, normal observations,

The standardized vector is obtained from the standardized values of X*i* (*i*=1, 2,*k*). MD measures the distance in multidimensional spaces by accounting for the correlation among the attributes. The statistical meaning of MD is the nearness of an unknown point to the mean of the group. The following is the formula used to calculate MDs:

$$MD_j = D_j^2 = \frac{1}{k} Z_{ij}^T C^{-1} Z_{ij} \qquad (2)$$

Where C^{-1} is the inverse of the correlation matrix and T is the transpose of the standard vector. According to Taguchi, the average value of the MDs is 1 for all the observations in the MS. For this reason, MS is also called the unit space. [2]

The second step is to validate the measurement scale. In order to validate the measurement scale, observations outside of MS are used, usually abnormal observations. The mean value, standard deviation and correlation matrix of normal observations are used to calculate the MD of the abnormal observations. For good measurement scales, the MDs of the abnormal observations are larger than the MDs of the normal observations.

The third step of MTS is to optimize the system. For this purpose, Taguchi's orthogonal array (OA) and signal to noise array (SN) are very useful to identify which attributes are important. In the experiment, every factor is assigned to a column in OA, and every row represents the experiment combination of a run. A two level OA is used to represent inclusive and exclusive. Each attribute will be used or neglected with respect to the OA and the SN ratio is calculated.

There are many different types of SN ratio; however, MTS uses the dynamic SN ratio. In an ideal function, the output is equal to the signal. The SN ratio evaluates the quality of measurement. It reflects the severity of the abnormalities and the difference of the average SN values of each attribute when it included and excluded. The classification ability is compared with feed forward artificial neural network. In the aspect of data size, efficiency and time, MTS show good performance compare to neural network. Equation (3) shows the dynamic SN ratio.

$$SN = -10 log\left(\frac{1}{t}\sum_{j=1}^{t}\frac{1}{MD_j^2}\right) \tag{3}$$

For a given attribute Xi, SN+ represents the average SN ratio of including the attribute Xi. SN- represents when Xi is excluded.

$$Gain = SN^+ - SN^- \tag{4}$$

If the gain is positive, the attribute is use, if not it is neglected. After the confirmation test, the optimization results are compared with the before and after.

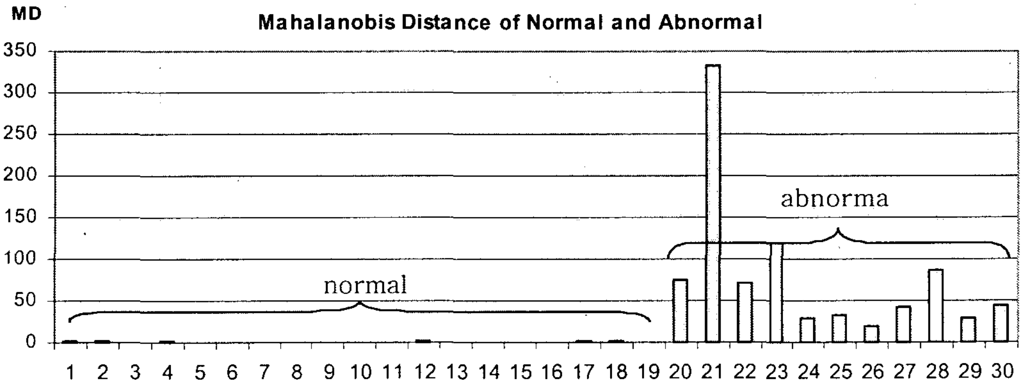
4.Simulation

The study used the breast cancer data from the UCI machine learning repository, which was collected at the University of Wisconsin by W. H. Wolberg. The goal is to predict whether a tissue sample taken from a patient's breast is malignant or benign. There are a classe, nine numerical attributes, and 699 observations. Sixteen instances contain a single missing attribute value and are removed from the analysis.

<Table 1> Wisconsin Breast Cancer Attributes Name and Data Type

Attribute	Domain
1. Clump Thickness	1-10
2. Uniformity of Cell Size	1-10
3. Uniformity of Cell Shape	1-10
4. Marginal Adhesion	1-10
5. Single Epithelial Cell Size	1-10
6. Bare Nuclei	1-10
7. Bland Chromatin	1-10
8. Normal Nucleoli	1-10
9. Mitoses	1-10
10. Class	2 for Benign, 4 for Malignant

Using a Matlab random number generator, 30 data sets are selected from entire 683 data sets. The data separated by benign and malignant to use as a normal (healthy) and abnormal (unhealthy) data. Nineteen data sets are benign and eleven data sets are malignant. The healthy (benign) data sets are generalized by equation (1). The correlation matrix and inverse of correlation matrix are calculated. The Mahalanobis distances of the selected data sets are calculated by equation (2). Fig. 1 shows the magnitude of normal MD is very small compare to those of abnormal MD. That means the selected data sets have ability of discrimination of healthy and unhealthy data. Table 2 shows the attribute data and calculated Mahalanobis distance. In the case of the normal (healthy) data sets, the MD value is very small and the average of MD is close to one. The abnormal MD values are larger than normal which illustrates the classification ability of MD. Figure 1. represents the MD of the normal and abnormal data. With previous results, the selected data can classify the normal and abnormal case.



<Fig. 1> Normal and Abnormal MD

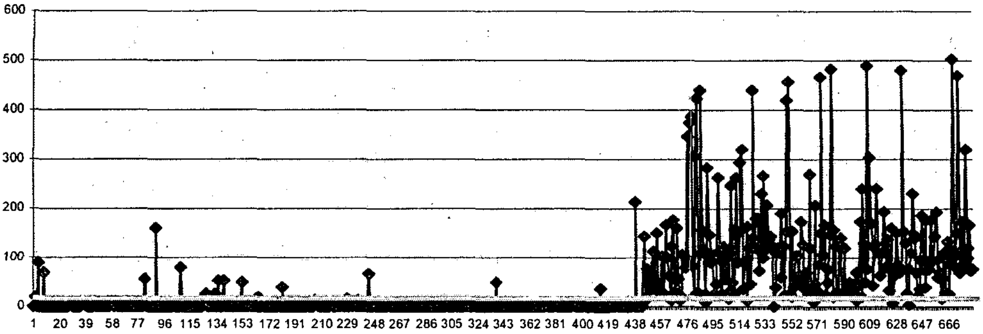
	A	B	C	D	E	F	G	H	I	Class	MD
1	5	1	1	3	4	1	3	2	1	2	1.754
2	5	1	2	10	4	5	2	1	1	2	1.608
3	1	1	1	1	2	1	2	1	1	2	0.572
4	1	1	1	1	2	5	1	1	1	2	1.635
5	5	1	1	6	3	1	1	1	1	2	0.847
6	5	1	1	1	2	1	2	2	1	2	0.844
7	3	1	1	1	2	1	3	1	1	2	0.395
8	4	1	2	1	2	1	3	1	1	2	0.444
9	5	1	1	1	2	1	1	1	1	2	0.383
10	5	1	1	1	2	2	2	1	1	2	0.941
11	4	1	3	3	2	1	1	1	1	2	1.895
12	5	2	2	2	2	1	1	1	2	2	0.436
13	3	1	1	3	2	1	1	1	1	2	0.436
14	5	1	3	1	2	1	2	1	1	2	0.888
15	5	1	1	1	2	1	2	2	1	2	0.844
16	1	1	1	2	2	1	3	1	1	2	0.855
17	1	3	1	1	2	1	2	2	1	2	1.681
18	4	2	1	1	2	2	3	1	1	2	1.039
19	5	1	1	1	2	1	1	1	1	2	0.470
20	10	10	8	10	6	5	10	3	1	4	75.8
21	10	10	10	7	9	10	7	10	10	4	333
22	7	9	4	10	10	3	5	3	3	4	71.82
23	5	10	10	8	5	5	7	10	1	4	119.5
24	5	5	5	2	5	10	4	3	1	4	28.76
25	8	6	5	4	3	10	6	1	1	4	33.5
26	8	4	4	1	2	9	3	3	1	4	19.76
27	4	2	3	5	3	8	7	6	1	4	43.02
28	6	1	3	1	4	5	5	10	1	4	85.88
29	10	4	7	2	2	8	6	1	1	4	28.15
30	9	5	8	1	2	3	2	1	5	4	43.34

<Table 2> Selected Wisconsin Breast Cancer Data and Mahalanobis Distance

Next, with the correlation matrix, standard deviation and mean of healthy data sets, try to diagnosis of entire data and check the accuracy. Fig 3. show the magnitude of MD with respect to benign and malignant patients.

There are two types of error in diagnosis. Type 1 error, false positive, mis-judged normal as abnormal, and type 2 error, false negative, judging abnormal as normal. The threshold value can be decided by reducing the total loss caused by the two types of error instead of paying attention to only one.

The MS construct with only 30 data sets out of the 683 total data sets. From the pre-constructed MS, the MD of each data sets was calculated and compare the magnitude of MD with threshold value. After testing whole data sets, diagnosis accuracy is 95.9% with 19 type 1 error and type 2 error.



<Fig. 2> Wisconsin Breast Cancer Test Results

Finally, using OA and SN ratio, the significance of each attribute is tested. With an L12 orthogonal array, the levels of each row decide whether that attribute is used or unused. Since the L12 orthogonal array has 12 rows, 12 different cases exist and the Mahalanobis space of each case is calculated. Abnormal test data is used for calculating the SN ratio. Since the deviation of the abnormal case is bigger than those of normal case, the difference between using some attributes and excluding attributes can be easily detected. Table 3 represents the L12 orthogonal array and SN ratio.

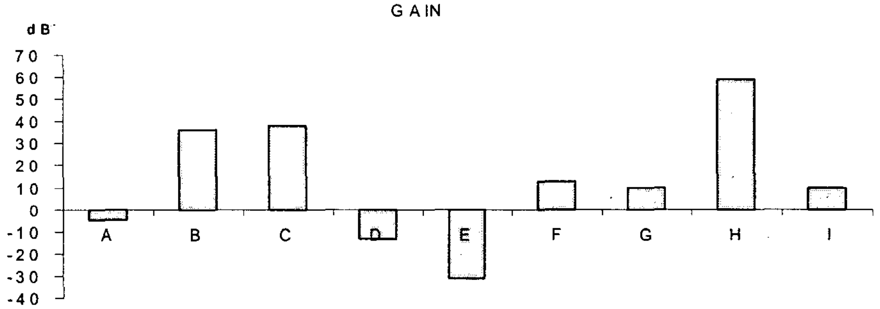
<Table 3> L12 Orthogonal Array and SN Ratio

	A	B	C	D	E	F	G	H	I			S/N Ratio
1	1	1	1	1	1	1	1	1	1	1	1	31.6
2	1	1	1	1	1	2	2	2	2	2	2	16.45
3	1	1	2	2	2	1	1	1	2	2	2	31.12
4	1	2	1	2	2	1	2	2	1	1	2	20.11
5	1	2	2	1	2	2	1	2	1	2	1	13.9
6	1	2	2	2	1	2	2	1	2	1	1	16.07
7	2	1	2	2	1	1	2	2	1	2	1	18.27
8	2	1	2	1	2	2	2	1	1	1	2	26.06
9	2	1	1	2	2	2	1	2	2	1	1	26.18
10	2	2	2	1	1	1	1	2	2	1	2	7.24
11	2	2	1	2	1	2	1	1	1	2	2	26.52
12	2	2	1	1	2	1	2	1	2	2	1	29.79

<Table 4> Level Average SN Ratio and Gain

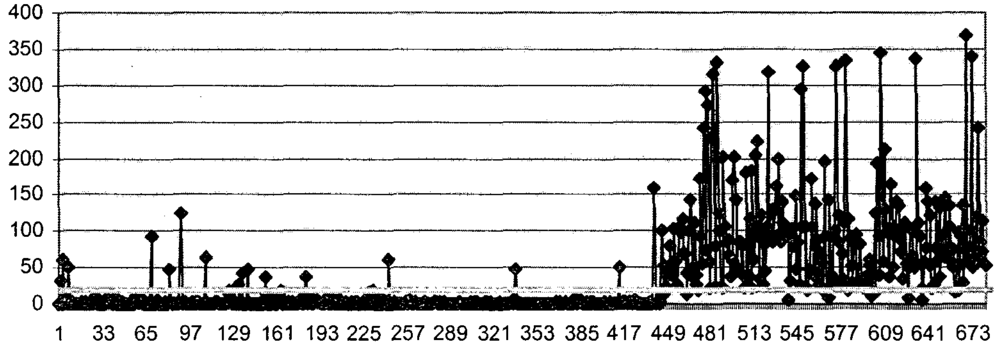
	A	B	C	D	E	F	G	H	I
Used	129.3	149.7	150.7	125.0	116.2	138.1	136.6	161.2	136.5
Unused	134.1	113.6	112.7	138.3	147.2	125.2	126.7	102.2	126.9

Gain	-4.8	36.04	38	-13.2	-31	12.96	9.82	59	9.6



<Figure 3> Gain of attributes

Table 4. and Figure 3. show the optimization results. The negative gained attributes are less affects to the system and can be neglected for the next process. The attributes B, C, F, G, H and I are selected and construct another MS and calculate MD. Figure 4 show the classification results. With only 6 attributes used for this test and the accuracy of diagnosis is 95.61% with 19 type 1 error and 11 type 2 error.



<Fig. 4> Optimization results (6 attributes)

5.Conclusion

The main idea of Taguchi's quality engineering is analyzing complicated and time consuming multivariate system with efficiency. In multivariate system, the existence of multicollinearity and partial correlations makes it difficult to analyze the system. But in the most cases, neglecting that is a trade off between time and effort to analyzing a system. In this study, for diagnosis of breast cancer, MTS shows good diagnosis ability with the accuracy of 95.9%. After optimization process, we can reduce three attributes for diagnosis but the difference of diagnosis accuracy is not significantly diminished (95.61%).

However, MTS have some points to be improved. In MTS, it is very important to select a "normal" or "healthy" group before constructing the MS and discrimination ability largely depends on how well select the normal samples. Determination of the threshold value and application for multiple class cases are areas for further study.

References

[1] Taguchi, G and R. Jugulum, "New Trends in Multivariate Diagnosis", *Indian Journal of Statistics*, 62, Series B, 2 233-248 (2000).

[2] Taguchi, G and R. Jugulum, The Mahalanobis-Taguchi Strategy: A Pattern Technology System, John Wiley & Sons, Inc., 2002.

[3] Jain, Anil K., Duin, Robert P.W., Mao, Jianchang, "Statistical Pattern Recognition: A Review", *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 1, Jan. 2000.

[4] Taguchi, S., "Mahalanobis Taguchi System", *ASI Taguchi Symposium*, 2000.

[5] Lande, U., "Mahalanobis Distance: A Theoretical and Practical Approach", http://biologi.uio.no/fellesavdelinger/finse/spatialstats/Mahalanobis%20distance.ppt, 2003.

[6] Hayashi, S., Y. Tanaka, and E. Kodama, "A New Manufacturing Control System using Mahalanobis Distance for Maximizing Productivity", *IEEE Transactions*, 59-62, 2001.

[7] Asada, M., "Wafer Yield Prediction by the Mahalanobis-Taguchi System", *IIE Transactions*, 25-28, 2001.

[8] Wu, Y., "Pattern Recognition using Mahalanobis Distance", *TPD Symposium*, 1-14, 1996.

[9] Taguchi G., S. Chowdury, and Wu Y., The Mahalanobis Taguchi System, McGraw Hill Press New York, 2001.

[10] Hassoun, M.H., "Fundamentals of Artificial Neural Networks", Cambridge, MA, MIT Press, 1995.