International Journal of Innovative
Computing, Information and Control
Volume **18**, Number **1**, February **2022**

ICIC International ©2022 ISSN 1349-4198
pp. **173–181**

FEATURE ANALYSIS USING MAHALANOBIS-TAGUCHI METHOD AND GENETIC ALGORITHM FOR RECORDED TV DATA

Shinichi Murata* and Hiroshi Morita

Graduate School of Information Science and Technology
Osaka University
1-5 Yamadaoka, Suita, Osaka 565-0871, Japan
morita@ist.osaka-u.ac.jp
*Corresponding author: shinichi.murata@ist.osaka-u.ac.jp

Received July 2021; revised October 2021

Abstract. *Data analysis has been attempted by numerous companies in a variety of industries, but it is often unsuccessful either because a complete dataset is unavailable or because the data are complicated and cannot be analyzed without special expertise. Moreover, it is often difficult to explain the results of complex AI and machine learning algorithms to those charged with acting on the results. To address these problems, we propose a data analysis approach that combines the Mahalanobis-Taguchi (MT) method with a stochastic optimization model to automatically identify data features (variables/ attributes) that directly affect the analysis. The proposed approach can be used to analyze even small datasets and eliminates tasks that require higher-level knowledge and skills (such as feature selection) when the data are complicated. Importantly, the method automates the analysis and improves the explainability of the analytical results. The effectiveness of the proposed approach is verified using recorded TV data.*
Keywords: Mahalanobis-Taguchi method, Genetic algorithm, Feature selection, Data analysis, Recorded TV data

1. **Introduction.** As ICT technology continues to evolve, the environment for utilizing IoT, AI, and big data is becoming more and more familiar. Increasingly, companies are choosing to use data analysis services provided by such entities as Amazon and Microsoft for their data analysis work. However, many companies are unable to effectively analyze the big data acquired by IoT using AI and thus fail to realize its full potential.

One of the problems is that companies often lack sufficient datasets, which can mean, for example, not having enough teacher data or having only a single class of data with which to work. In many cases, a company is able to obtain only partial data due either to the division of the organization that manages the data or to various design defects. As a result, the datasets they hold are incomplete and violate the basic assumptions of effective data analysis. When using common data analysis services, it takes time and effort to prepare the data. Another important challenge is that the data features can be complex in nature and enormous in number. With the collection of diverse data that include elements such as purchase behavior, purchase history, and various types of operation logs, the characteristics of the data become more complex and the number of features increases significantly. Without a deep understanding of the data, it is difficult to select and analyze those features that are most relevant to a project's specific purpose from among the huge number of features that may be available. In addition, analysis using AI and machine learning may be rejected because the results are difficult to explain to stakeholders given the large number of black box elements involved.

DOI: 10.24507/ijicic.18.01.173

In research related to these issues, ways to analyze data using only a small number of datasets, such as datasets with only one class, are being studied in the field of anomaly detection. Methods such as the k-Nearest Neighbor (k-NN) method [1], Local Outlier Factor (LOF) [2], and One-Class SVN (OCSVN) [3] are typical examples. LOF and OCSVN are particularly effective in detecting outliers in data with complex distributions. However, when applied to actual problems, both have complicated characteristics, such as the need for sophisticated and precise threshold- and parameter-setting in accordance with the data being analyzed. Feature selection methods have been studied mainly in the fields of machine learning, deep learning, and AI for the purpose of avoiding noise pattern learning and improving learning speed. In addition to the conventional forward selection and backward elimination approaches, efforts are being made to improve analysis [4,5] performance by implementing SVM and k-NN using optimization methods for feature selection. The explainability of AI and machine learning has become more important in recent years, and research on explainable AI is currently underway. Permutation importance [6] is a method that does not depend on the opportunity learning algorithm. In this approach, it is possible to show which features contribute significantly to the model generated from various algorithms; however, it is not possible at present to explain precisely why the features were selected.

In business, while the need for data analysis using AI and machine learning is rapidly expanding, there is a shortage of data scientists. As a consequence, many business organizations find it difficult to meet the challenges of selecting the proper data analysis algorithm, tuning the appropriate parameters, identifying the optimal number of features, or choosing the best method for evaluating the selected features to fit their specific needs.

In view of these circumstances, this paper proposes a feature analysis method that can be implemented by individuals having no special expertise in data analysis and that allows pertinent academic research to be directly applied in a practical business setting. The proposed method eliminates the need for advanced expertise of the type required for choosing a data analysis algorithm, setting parameters, selecting features, and evaluating results. With this method, feature analysis is performed using the Mahalanobis-Taguchi (MT) method [7,8], which enables analysis using a small dataset involving only one class of data and does not require complicated parameter-setting. In the basic MT method, feature selection/evaluation employs an orthogonal array; however, this step is difficult and requires a level of specialization. In order to minimize the complexity of feature selection and ensure explainability, the proposed approach applies optimization methods used in machine learning to the MT method. This allows for the automatic selection of features and a visualization (evaluation) of the contribution of the features selected.

The remainder of the paper proceeds as follows. An outline of the proposed method is provided in Section 2. Section 3 describes the results of applying the proposed method to recorded TV data and verifying its effectiveness. Section 4 presents conclusions and discusses future prospects.

2. **Outline of the Proposed Method.**

2.1. **Feature analysis by applying the MT method and the genetic algorithm optimization method.**

In this paper, we match the characteristics of a data group (A) that belongs to a certain space with data group (B) whose space is unknown. Data group (A) is then automatically extracted from the data group whose space is unknown.

For the group of data known to belong to a certain space (A), the features to be used are first automatically determined using a genetic algorithm [9,10], one of the stochastic optimization methods used in Operations Research (OR). Next, the MT method is applied

to the selected features to form a unit space that serves as a reference for judgment based on the Mahalanobis distance; model evaluation is then performed based on the selected features using the test data. The cycle of feature selection, unit space creation, and model evaluation is then repeated, and the feature to be finally used is determined based on the evaluation results.

The Mahalanobis distance is then calculated by the MT method for the data group whose space is unknown, and the distance from the unit space is determined. As a result, we can construct a mechanism to extract the data belonging to a certain space (A) from the data group whose space is unknown. Figure 1 describes the analysis process.

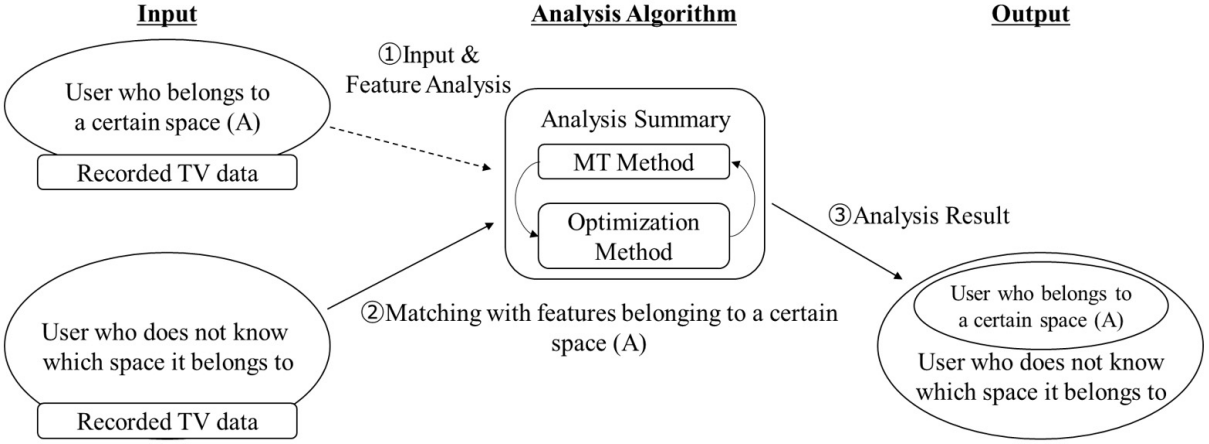


FIGURE 1. Image of analytical processing

2.2. **Outline and application of the MT method.** The MT method [11] is a method for determining outlier values based on the Mahalanobis distance. Although Mahalanobis distance-based outlier determination has been a long-time staple in the world of statistics, Taguchi et al. [16] added a number of useful elements and developed many simple methods around it. Although there are no prior studies applying the MT method to the extraction of features for recorded TV data, extensive MT-related research has been conducted in quality engineering [12] and anomaly detection [13,14]. Figure 2 describes the judgment procedure used in the MT method.

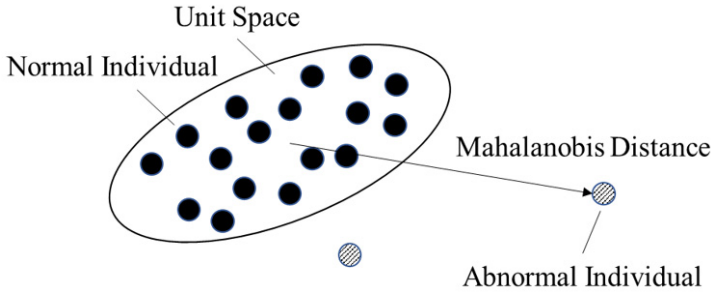


FIGURE 2. Judgment by MT method

In the MT method, the analysis is performed according to the following procedure.

1) Unit space (normal) data are determined and standardized.

2) The correlation coefficient matrix and the inverse matrix are calculated from the standardized data.

3) The Mahalanobis distance is calculated from the standardized data to determine the threshold.

4) Item selection is performed using an orthogonal array from the abnormal data.

The Mahalanobis distance is calculated for the new data using the selected items, and a judgment is made based on a pre-set threshold value.

The mean $\boldsymbol{\mu} = \{\mu_1, \mu_2, \ldots, \mu_m\}$ and standard deviation $\boldsymbol{\sigma} = \{\sigma_1, \sigma_2, \ldots, \sigma_m\}$ of n data values of $\boldsymbol{X} = \{X_{m1}, X_{m2}, \ldots, X_{mn}\}$ of m dimension belonging to the unit space are calculated and the standardized $\boldsymbol{X}$ is produced according to Equation (1):

$$\hat{X}_{ij} = \frac{X_{ij} - \mu_i}{\sigma_i} \tag{1}$$

Let the element $a_{i1,i2}$ of correlation coefficient matrix $\boldsymbol{A}$ be defined by Equation (2), and let each element of the inverse matrix of $\boldsymbol{A}$ be $\bar{a}_{i1,i2}$.

$$a_{i1,i2} = \frac{1}{n}\sum_{j=1}^{n} X_{i1j} X_{i2j} \tag{2}$$

The square of the Mahalanobis generalized distance MD_Y of the m-dimensional observation data $\boldsymbol{Y}$ is given by Equation (3). By setting a threshold value for this MD, it is possible to determine whether an observed data value is normal or abnormal.

$$MD_Y^2 = \frac{1}{m}\sum_{ij} \bar{a}_{ij} \left(\frac{Y_i - \mu_i}{\sigma_i}\right)\left(\frac{Y_j - \mu_j}{\sigma_j}\right) \tag{3}$$

Next, according to the orthogonal array of the two-level system, the items used for the calculation of l abnormal data values $\boldsymbol{Z} = \{Z_1, Z_2, \ldots, Z_l\}$ are limited, and the recalculated Mahalanobis generalized distance is set as MD_{Z_i}. The SN ratio η [db] is defined by Equation (4):

$$\eta = -10 \log \frac{1}{l}\left(\frac{1}{MD_{Z_1}^2} + \frac{1}{MD_{Z_2}^2} + \cdots + \frac{1}{MD_{Z_l}^2}\right) \tag{4}$$

Since this is the signal-to-noise ratio of the larger-is-better response, η shows a higher evaluation value as MD increases. According to the orthogonal array of the two-level system, the average of the SN ratio when a certain item is used is $\bar{\eta}_a$, and the average of the SN ratio when it is not used is $\bar{\eta}_b$. Then, the significance e_i of the item is calculated by Equation (5):

$$e_i = \bar{\eta}_a - \bar{\eta}_b \tag{5}$$

By setting a threshold value for the significance e_i, the calculation items can be optimized and the cause of the abnormality can be identified.

In this paper, for item selection (feature selection) and evaluation using the MT method, item selection is performed at the same time that the unit space is determined by applying an optimization method rather than using an orthogonal array. The Mahalanobis distance is then calculated and the model is evaluated using the selected items. So long as the evaluation in the orthogonal array is precisely designed and evaluated, the features that have a great influence on the relationship and the results can be identified with a small amount of calculation. On the other hand, this often involves a substantial amount of manual work offline, and when there are many features of the data to be applied, the time to design the evaluation by orthogonal array and the evaluation according to the design can be significant. In addition, although design of experiments and parameter design methods have been proposed for evaluation design and evaluation, specialized knowledge and experience are required, and it is difficult for everyone involved to produce work of equal quality. In addition, item selection using an orthogonal array is performed

using abnormal data rather than normal data. However, as noted earlier, in 'real-world' companies, normal data and abnormal data are often not available in advance, and in many cases they cannot be applied.

Using an optimization method for item selection (feature selection) ensures an accuracy and quality that is practically difficult to achieve by offline work, particularly when a large number of features, as in our recorded TV data, are involved. Such an approach makes it possible to automatically select and evaluate items while maintaining these essential properties. It is also advantageous when deep knowledge of the data is required and the design of the orthogonal array can vary, or when the data that can be prepared in advance are strictly normal data.

2.3. **Outline of the genetic algorithm.** As noted above, a genetic algorithm is used as the optimization method. Genetic algorithms have been applied in various fields, mainly in OR, and are easy to apply to a variety of problems due to their simplicity and uniformity.

The genetic algorithm is a type of metaheuristic algorithm proposed by Holland [17] of the University of Michigan in 1975 to search for an approximate solution. Considering the evolutionary process of living organisms, individuals with high environmental adaptability in the population have a high probability of surviving and leading the next generation. In modeling this principle, the most adaptable individual can be identified. Although there is no guarantee that the method will always find the optimum solution, it typically finds approximate solutions with a relatively high degree of accuracy in a short time. The method has been applied to a variety of problems, including combinatorial optimization problems and NP-hard problems [15].

The algorithm consists of 10 steps, with step 9 serving as an auxiliary step for speeding up the convergence that is used only when the weights of each vertex are relatively balanced. Briefly, the 10 steps include the following.

Step 1. Initial setting: Set the value of α to a predetermined value.

Step 2. Generation of the initial population: Create an initial population. The initial population is composed of many individuals, and each individual presents a proposal for division. Random values are set for the genes that make up the chromosome in order to give the initial population a random division plan.

Step 3. Sorting (in ascending order):

1) Sort the gene values in ascending order for each individual.

2) If the gene values (values for determining the group) are the same, the vertex numbers should be in ascending order.

Step 4. Positioning when dividing into groups: Let S_r be the r-th vertex after sorting; the sum of the weights of the vertices so far is determined according to Equation (6):

$$PW_r = \sum_{i=S_1}^{S_r} w_i \tag{6}$$

Here, r indicates the boundary where vertex S_r belongs to group NB and the next vertex S_{r+1} belongs to group $NB + 1$ obtained by the following equation:

$$|(W/D) \times NB - PW_r| \to minimum \tag{7}$$

Step 5. Calculation of fitness of each individual: The sum of the weights of the sides across the groups (cut weights) is calculated. The smaller the value, the better the division.

Step 6. Selection: For the individuals that will remain in the next generation, some will be determined by the elite strategy and the rest by roulette selection. In the elite

strategy, those with high fitness are left unconditionally. In roulette selection, the ratio of each individual to the roulette board is established based on fitness; the roulette is turned to leave the hit individual.

Step 7. Crossover: Two individuals cross chromosomes to create a new individual. The two parents are selected by roulette selection. The parent chromosomes are uniformly crossed (a random number is taken for each gene to determine which parent inherits it), and a chromosome (individual) is created as a child.

Step 8. Mutation: One of the genetic manipulations is performed on an individual. Part of the genetic information is changed with a predetermined probability. This provides an opportunity to escape the local solution.

Step 9. Maximum number of cuts vertex detection: The vertex with the maximum sum of the weights of the sides that span other groups is detected for each group. This is stochastically transferred to another group.

Step 10. Population evaluation: The procedure ends when there is no change in a certain generation or when the specified number of generations is reached. If this condition is not met, return to Step 3.

3. Application and Verification of the Proposed Method for Recorded TV Data.

3.1. **Verification overview.** To verify the effectiveness of the proposed method, we used recorded TV data, which has a large number of data values and for which it is relatively easy to express features. The recorded TV data used in this experiment consist of recorded program content, broadcast date and time, channel information, meta information such as genre information, viewing presence/absence, recording mode, and information such as which device was recorded. In the verification, we used 892 data points and 25 features. Table 1 describes some of the recorded TV data.

Table 1. Recorded TV data example (partial excerpt)

Item	Supplementary explanation
Recording device	
Number of recordings	• Uses 25 features
Channel information	• Targets recorded TV data included in 892 recording devices
...	

The sequence of steps proceeded as follows. First, the recorded TV data group of the user belonging to a certain space is used as an input, and the features used for analysis are automatically selected at random. The MT method is then applied using the selected features to generate a unit space (with the calculation of the Mahalanobis distance), and the model is evaluated using the selected features. Test data are used to evaluate the model, and the evaluation is performed based on the correct answer rate of the space specific to the test data. Next, using the genetic algorithm, the cycle of feature selection, unit space generation, and model evaluation using the selected features is repeatedly performed to calculate the optimum features. Following this, the Mahalanobis distance of the recorded TV data of an uncertain user (i.e., a user whose membership in the corresponding space is uncertain) is calculated by using the recorded TV data of the user as an input and employing the optimum feature amount for each space calculated in advance. It is then judged whether this user belongs to the corresponding space. Figure 3 shows the flow of the proposed method.

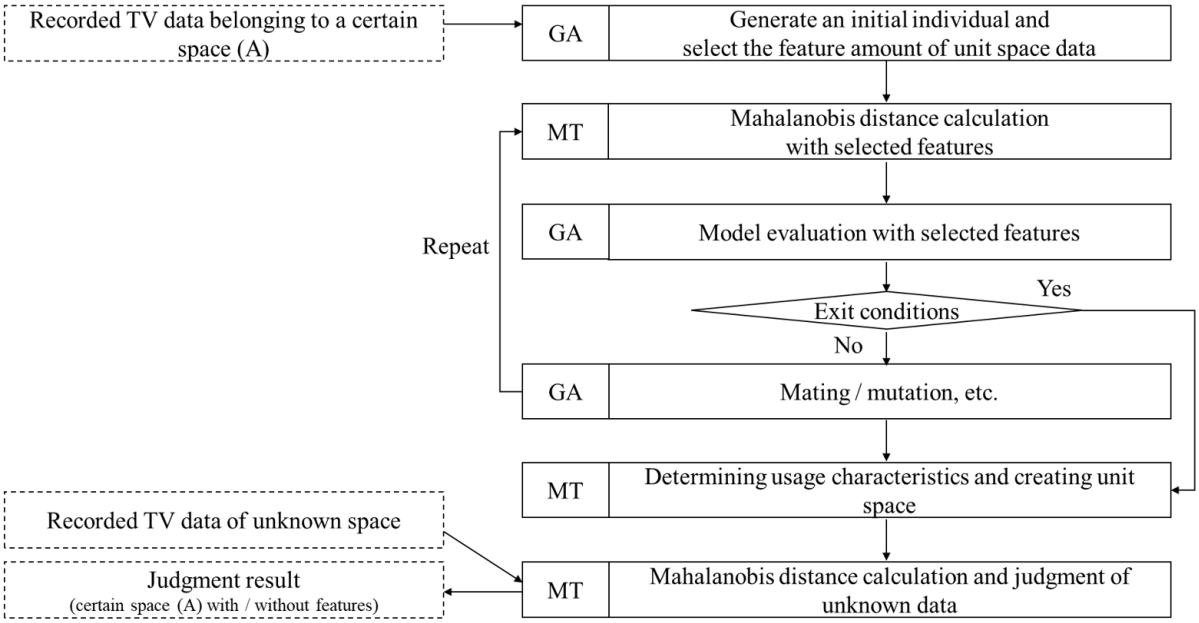


FIGURE 3. Flow of the proposed method

3.2. **Validity verification of the proposed method with 1-class data.** It is known that the reference data for creating the unit space belong to a certain space (A). When it is not known which space the verification target data belong to, the reference data are used to identify the data belonging to the certain space (A) from the verification target data. For this verification, 200 data values belonging to the certain space (A) were used as the reference data for creating the unit space. As verification data, 140 data values belonging to the certain space (A) and 752 data values whose space was unknown were used. The Mahalanobis distance used as the threshold value for judgment was set to 4, a numerical value often used in quality engineering. For each parameter of the genetic algorithm, 100 individuals, 10 generations, 0.2 mutations, and 0.1 elite selections were used. In addition, in order to confirm the effectiveness of the proposed method for the recorded TV data, we compared its performance with that of two other methods: One-Class Support Vector Machine (OCSVM) and Local Outlier Factor (LOF).

Table 2 and Figure 4 show the results for the TV test data. As can be seen here, the correct answer rate for the proposed method was substantially higher than for the other methods, confirming, in this case, the effectiveness of our method for the recorded TV data.

TABLE 2. Correct answer rate for test data

Method	Correct answer/ overall data	Correct answer/ Space (A) data	Correct answer/ unknown space data
Proposed method	870/892	136/140	734/752
OCSVM	752/892	10/140	742/752
LOF	753/892	21/140	732/752

4. **Conclusions and Future Work.** Based on the recorded TV data of users belonging to a certain space, we determined the unit space by applying the MT method and a stochastic optimization model. Data for users belonging to a certain space were then extracted from the data for users whose space was uncertain. Importantly, the proposed

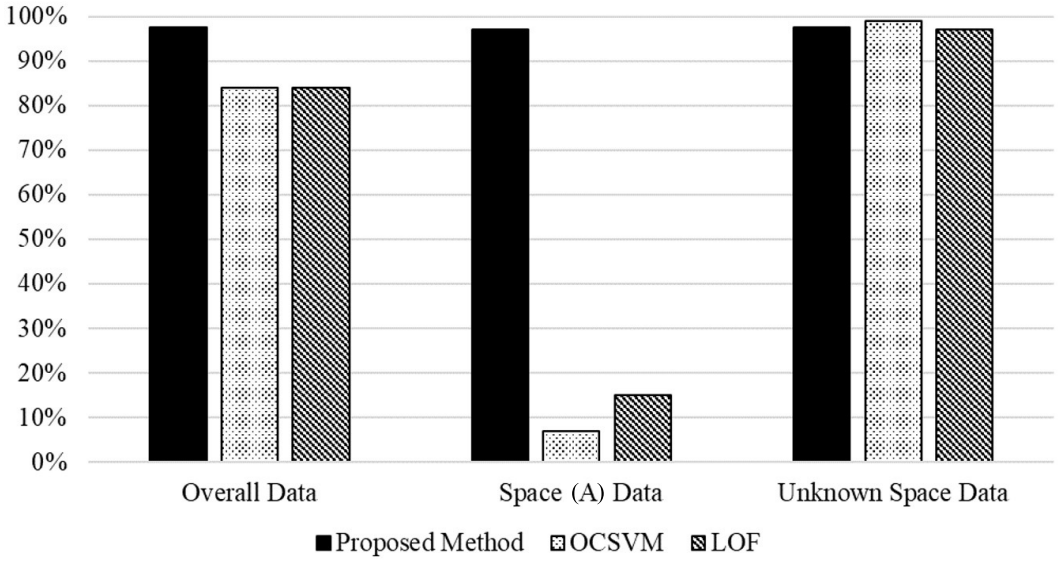


FIGURE 4. Judgment results for test data

method can automate those parts of a data analysis that require expert knowledge and special skills, such as feature selection. Since the proposed approach makes it possible to confirm the features that were used to produce the result, the method is easy to apply to the practical data analyses in which a company is likely to engage. As future work, we intend to develop data analysis logic corresponding to more varied data types by making the model more generalized.

REFERENCES

[1] T. Cover and P. Hart, Nearest neighbor pattern classification, *IEEE Trans. Information Theory*, vol.13, no.1, pp.21-27, 1967.

[2] M. M. Breunig, H.-P. Kriegel, R. T. Ng and J. Sander, LOF: Identifying density-based local outliers, *ACM SIGMOD Record*, vol.29, no.2, pp.93-104, 2000.

[3] Y. Chen, X. S. Zhou and T. S. Huang, One-class SVM for learning in image retrieval, *Proc. of 2001 International Conference on Image Processing (Cat. No. 01CH37205)*, vol.1, pp.34-37, DOI: 10.1109/ICIP.2001.958946, 2001.

[4] J. Sakhnini, H. Karimipour and A. Dehghantanha, Smart grid cyber attacks detection using supervised learning and heuristic feature selection, *2019 IEEE 7th International Conference on Smart Energy Grid Engineering (SEGE)*, pp.108-112, 2019.

[5] T. Ahmad and M. N. Aziz, Data preprocessing and feature selection for machine learning intrusion detection systems, *ICIC Express Letters*, vol.13, no.2, pp.93-101, 2019.

[6] A. Fisher, R. Cynthia and D. Francesca, Model class reliance: Variable importance measures for any machine learning model class, from the "Rashomon" perspective, *arXiv Preprint*, arXiv: 180101489, 2018.

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

[8] G. Taguchi, *Technology Development in MT System*, Japanese Standards Association, 2002 (in Japanese).

[9] D. Goldberg, *Genetic Algorithm in Search, Optimization, and Machine Learning*, Addison-Wesley, 1989.

[10] J. H. Holland, Escaping brittleness: The possibilities of general-purpose learning algorithms applied to parallel rule-based systems, *Machine Learning*, pp.593-623, 1986.

[11] Y. Nagata, Several properties of MT system and improved procedures, *Journal of Applied Statistics*, vol.42, no.3, pp.93-119, 2013 (in Japanese).

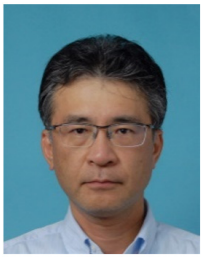
[12] A. Makabe, K. Takada and H. Yano, Application of mahalanobis distance for the automatic inspection of solder joints, *Journal of Quality Engineering Society*, vol.6, no.6, pp.66-73, 1998 (in Japanese).
[13] M. Takahama and N. Mikami, Detection of abnormal signs for gas turbine power plant, *Journal of Quality Engineering Forum*, vol.20, no.4, pp.437-443, 2012 (in Japanese).
[14] M. Ohkubo and Y. Nagata, Anomaly detection in high-dimensional data with the Mahalanobis-Taguchi system, *Total Quality Management & Business Excellence*, vol.29, nos.9-10, pp.1213-1227, 2018.
[15] M. Yamamura, T. Ono and S. Kobayashi, Character-preserving genetic algorithms for traveling salesman problem, *Journal of the Japan Society of Artificial Intelligence*, vol.7, no.6, pp.1049-1059, 1992 (in Japanese).
[16] G. Taguchi, S. Chowdhury and Y. Wu, *The Mahalanobis-Taguchi System*, McGraw-Hill, 2000.
[17] J. H. Holland, *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence*, The University of Michigan Press, 1975.

Author Biography

Shinichi Murata received the Master degree in information science and technology from Osaka University in 2007.

Mr. Murata is currently Ph.D. student at the Graduate School of Information Science and Technology, Osaka University, Japan since 2018.

His main research interests include system optimization, mathematical modeling, machine learning and data analysis using real data.



Hiroshi Morita received the Ph.D. degree in engineering from Kyoto University in 1992; the Master degree in engineering from Osaka University in 1985. He worked at Osaka Prefectural University, Osaka City University and Kobe University.

Prof. Morita is currently a full-time professor at the Graduate School of Information Science and Technology, Osaka University, Japan.

His research interests include system optimization, mathematical modeling, statistical data analysis, and quality management.