Feature Selection in Big Data by Using the enhancement of Mahalanobis–Taguchi System; Case Study, Identifiying Bad Credit clients of a Private Bank of Islamic Republic of Iran

Shahin Ordikhani[1*], Sara Habibi[2]
[1]Industrial Engineering, Industrial and Mechanical Engineering, Azad University, Qazvin, Iran
[2]Urmia University
*Email of Corresponding Author: sh.shahin2011@gmail.com

Received: May 8, 2019; Accepted: July 15, 2019

Abstract
The Mahalanobis-Taguchi System (MTS) is a relatively new collection of methods proposed for diagnosis and forecasting using multivariate data. It consists of two main parts: Part 1, the selection of useful variables in order to reduce the complexity of multi-dimensional systems and part 2, diagnosis and prediction, which are used to predict the abnormal group according to the remaining useful variables. The main purpose of this research is presenting a new method to select useful variables by using and combining the concept of Mahalanobis distance and Integer Programming. Due to the inaccuracy and the difficulties in selecting the useful variables by the design of experiments method, we have used an innovative and accurate method to solve the problem. The proposed model finds the solutions faster and has a better performance than other common methods.

Keyword
Multivariate Analysis System, Mahalanobis-Taguchi System, Mixed Integer Programming, Machine Learning

NOMENCLATURE

Indexes		Parameters	
I	Variable index *(i=1,...,n)*	MD_j	The Mahalanobis distance of j^{th} observation
J	Observation index *(j=1,...,n)*	S_i	Standard deviation of i^{th} variable
N	The number of observations	Z_{ij}	The normalized value of i^{th} variable from j^{th} observation
K	The number of variables	S_z	The standard deviation of the normalized values
Variables	Particle Specific density	C	The correlation matrix
x_i	Represents the variables of the multivariate system	C^{-1}	Inverse correlation matrix
X_{ij}	i^{th} variable from j^{th} observation	Mmd_N	The mean of MDs for normal group
		Mmd_{AN}	The mean of MDs for abnormal group
		Std_N	The standard deviation of MDs for normal group
		Std_{AN}	The standard deviation of MDs for abnormal group
		$Diffmed$	The MDs' mean difference between normal and abnormal groups
		r	The total number of observations in normal group
		ra	The total number of observations in abnormal group

1. Introduction

Analyzing the real world's problems and finding the optimum results have drawn researchers' attention for a long time. Since the real world's problems are affected by various variables that each of which affects the others, they create complex systems. In this regard, multi-variable systems analysis has been developed to deal with these kinds of complexities. From one point of view, increasing the number of variables reduces accuracy and increases complexity. From another point of view, analyzing multiple variables may not be possible or economical. In the last years, multivariate analysis systems have been presented. The main purpose of the multivariate analysis systems is to simplify the information and reducing the dimensions of the problem. There are different multivariate analysis systems such as: All subsets regression, Sequential search algorithm, Genetic algorithm (GA), Particle swarm optimization (PSO), ant colony optimization (ACO), Least Absolute Shrinkage and Selection Operator (LASSO), Variables Importance on PLS (Partial Least Squares) projections (VIP).All mentioned methods have their own defects. For instance, all subsets regression examines all possible subsets of the variables (2^p-1), which is reliable, but it has high computational complexity.The central premise when using a feature selection technique is that the data contains some features that are either redundant or irrelevant and as a result, be removed without incurring much loss of information. Redundant and irrelevant are two distinct notions, since one relevant feature may be redundant in the presence of another relevant feature with which it is strongly correlated. The sequential search algorithm does not examine all possible subsets, so, the optimal combination provided by it may not be optimal in reality. The Mahalanobis-Taguchi System (MTS) is a diagnostic and forecasting technique for multivariate data. MTS establishes a classifier by constructing a continuous measurement scale rather than learning from the training set directly. Therefore, it is expected that the construction of an MTS model will not be affected by the distribution of data. In the Mahalanobis Taguchi system, there are two kinds of misclassification. First, observations those are actually related to the normal group but are categorized in the abnormal group mistakenly. Second, observations those is actually related to the abnormal group but are categorized in the normal group mistakenly. These misclassifications will reduce the accuracy of prediction. In addition, as usually the variables are measured at the same time for each sample, it does not consider the correlation between variables, which reduces the accuracy of prediction. This investigation presents a new model that will increase the accuracy of prediction, reduce the solving period and computational complexity. In order to validation, results that are gained from the Taguchi method and the proposed model will be compared as a case study.

2. Literature Review

Feature Selection is the most important part to obtain the optimal subset in big data. Guyon and Elisseeff studied that the importance of the variables in terms of (a) the correlation output (y) and (b) disorder output data, depends on inputted variables (X) [1]. Alpaydin investigated that the maximum desirability of the model improves the accuracy of prediction on machine learning, but leads to excessive pairing and fails to generalize the model on non-machine learning [2]. Ng presented that it is important to minimize the number of variables for excessive pairing penalties and improve the accuracy of prediction for non-learning data [3]. In the Mahalanobis Taguchi System (MTS), feature selection is only done by improving the desirability of the MTS classification. Taguchi et al. have considered the maximization of the S/N[1] rate for feature selection [4]. A subset of variables is generated by using Taguchi's OAs[2] and the S/N rate of each subset is evaluated. Various studies have used S/N rates and OAs for selecting features in order to improve the accuracy of prediction of MTS classification in various applications. Lee and Teng have used a large number of financial rates as variables for making MTS classification that predicts important features of companies' bankruptcy by maximizing S/N rates [5]. Das and Datta have used MTS classification to predict the quality of hot rolled steel plates based on the chemical composition in which the important elements for the quality of steel plates are selected by maximizing the S/N rate [6]. Rai et al. developed an MTS classification for detecting the fractures of metal drilling tools by using S/N rates and OAs for predicting important signal vibrational parameters [7]. Yang and Cheng used the S/N rates and OAs to identify important inspection parameters for checking the quality of the flip chipsets and determining the quality of the produced chipsets by the MTS classification [8]. Abbasi et al. predicted damages for vehicle insurance in Iran's insurance companies and then compared their results with the neural network method [9]. At last, the results were in favor of the MTS. Jin and Chow developed the MD[3] based on the classification for the cooling fan system and induction motor [10]. They derived the important signal vibrational parameters based on the S/N rate and used them to create an MTS classification which identifies the errors in the induction motor. There are various studies in this regard; however, the S/N rate is calculated only by the abnormal observations' MD, regardless of the quality of prediction.

[1] 1- Signal to Noise
[2] 2- Orthogonal array
[3] 3- Mahalanobis Distance

Moreover, the OAs cannot generate all possible subsets of a variable. To address these problems, Pal and Maiti examined the feature selection problem as integer binary programming by minimizing theTWM[4] through BPSO[5] [11]. A case study in die casting quality prediction suggests that the new feature selection provides better classification accuracy than finding important variables by S/N rates and OAs. Resendiz and Roll-Flores used MTS classification to predict the quality of the car's pedals made by the injection molding process [12]. In order to improve the quality of prediction, variables are selected by Gompertz Binary Particle Swarm Optimization that reduces TWM. Resendiz at el predicted the quality of the truck's body assembly by using dimensional parameters. For this purpose, variables have been selected by binary Ant colony optimization, which led to TWM minimization. Barahimi and Aghaie used an integer mathematical programming model to select useful variables in a car insurance company and then predict the number of damages for car insurance [13]. They have improved the quality of prediction compared to other methods. Holcombe used TWM to optimize pattern recognition [14]. Li et al. also used this method to predict investment in an information company [15]. Okafor et al. studied the optimization of hardness strength response of plantain fibers reinforced polyester matrix composites by using Taguchi robust design [16]. They implemented a Taguchi robust design technique was for obtaining the highest S/N rate for the quality characteristics being investigated. Singh et al. are scrutinized the effect of main friction stir welding (FSW) parameters on the quality of AA 6063 plate welds [17]. They utilized the Taguchi approach in order to consider three levels, three factorial designs.

Although many studies have focused on the model's desirability by improving the accuracy of the prediction of MTS classification, few of them could set penalize for excessive pairing in the process of feature selection. The feature selection process in the proposed model is an objective function that improves the model's accuracy and avoids excessive pairing. An appropriate algorithm to determine the minimum subset of the features optimizes the objective function. The desirability of the model in the new feature selection method is also evaluated by the degree of dependence or conditional probability of the prediction class on the subset of the variables. The developed method includes a new criterion of proportionality and penalizes the excessive pairing. Then, the proposed model will be compared with other methods in the literature review for evaluating its accuracy. In general, there is a research gap in addressing the issue of excessive pairing during feature selection for the MTS classification.

3. Methodology

3.1 Review on the Mahalanbis Taguchi system

Taguchi and Jugulum believed that MTS has been developed in five steps: (1) Determining the number of observations in samples, (2) Creating the MS[6], (3) Evaluating the MS, (4) Identifying the useful variables, (5) Prediction and detection system [18].

[4] Total Weighted Misclassification

[5] Binary Particle Swarm Optimization

[6] Mahalanobis Space

3.1.1 Determining Number of Observations in the Sample

Taguchi and Jugulum opined that to obtain a correlation matrix, the number of samples should be more than the number of variables [18]. It is necessary, but may not be sufficient. Aman et al. believed that in the MTS, experimentally, the number of observations should be more than triple of the variables' number [19]. Foley investigated that for minimizing misclassification and figuring out whether the samples follow the same multivariate distribution or not, the ratio of the number of samples to the number of variables should be greater than three [20]. Reducing the number of samples increases misclassification. Leese et al. showed that when the number of samples in the normal group is small, these samples have a significant effect on the Mahalanobis space [21]. Moreover, adding or removing a new sample has a significant effect on the normal group.

3.1.2 Creating the MS

Fist, the observations should be classified into normal and abnormal groups. By defining the normal observations and calculating their MDs, the MS is formed. The MS contains the mean vector, standard deviation vector, and correlation matrix of the normal group. Mohan et al. proposed that the calculation of the MD to determine the MS consists of the following steps [22]:

Calculating the mean for each variable:

$$\bar{X}_i = \frac{\sum_{j=1}^{n} X_{ij}}{n} \tag{1}$$

Calculating standard deviation for each variable

$$S_i = \sqrt[n]{\frac{\sum_{j=1}^{n}\left(\bar{X}_i - X_{ij}\right)^2}{n-1}} \tag{2}$$

Normalizing for each variable

$$Z_{ij} = \frac{X_{ij} - \bar{X}_i}{S_i} \tag{3}$$

Confirming that the mean of normal data is equal to zero

$$\bar{Z}_i = \frac{\sum_{i=1}^{n} Z_{ij}}{n} = 0 \tag{4}$$

Confirmation that the standard deviation of the normalized data is equal to one.

$$S_z = \sqrt[n]{\frac{\sum_{i=1}^{n}\left(\bar{Z}_i - Z_{ij}\right)^2}{n-1}} = 1 \tag{5}$$

The correlation between the variables is calculated binary, then, the total correlation matrix (C) is formed. The correlation matrix between variable i and variable j is obtained as follows:

$$C_{ij} = \frac{\sum_{m=1}^{n}\left(Z_{im}.Z_{jm}\right)}{n-1} \tag{6}$$

Calculating the Mahalanobis Distance (MD)

$$MD_j = \frac{1}{k} \mathrm{Z_{ij}} C^{-1} Z_{ij}^{T} \tag{7}$$

3.1.3 Evaluating the MS

To evaluate the MS, the MD values for abnormal observations should be calculated. They will be normalized by mean, standard deviations and correlation matrix of the normal group. According to Taguchi and Jugulum, the MD values for abnormal observations should be greater than the MD values for normal observations [23]. It means that the normal observations are closer to the normal group's center than the abnormal observations.

3.1.4 Identifying the Useful Variables

In each sample of the Taguchi's orthogonal arrays table, variables can be allocated to level 1 or 2. Level 1 indicates the presence of that variable in the best subset and the level 2 indicates the absence of that variable in the best subset. According to Taguchi and Jugulum, for each row (each sample), S/N rate is obtained by using the MDs of abnormal observations as follows [23]:

$$\eta_q = -10 log \left[\frac{1}{t} \sum_{j=1}^{t} \frac{1}{MD_j} \right] \tag{8}$$

In which n_q and t stand for the S/N rate for q^{th} execution of orthogonal arrays and the number of abnormal observations under specified conditions respectively.

For each column (each variable), the Gain value is obtained as follows:

$$Gain = (Avg.\frac{S}{N} Ratio)_{Level1} - (Avg.\frac{S}{N} Ratio)_{Level2} \tag{9}$$

Variables with positive gain values are identified as useful variables and the rest of them are discarded. The ultimate Mahalanubis space, which has been derived from selected useful variables, can be used for subsequent diagnosis after determining the appropriate threshold value for binary categorization or directly for prioritizing based on the abnormality level. There are two major problems in the Taguchi orthogonal array table. First, based on Woodal et al. it cannot always provide the optimal subset of variables, and examining all possible subsets is a more guaranteed way, second, this method is useful when the variables do not affect each other [24].

3.1.5 Prediction and Detection System

The threshold value is used for predicting and detecting normality or abnormality of future observations. For this purpose, Su and Hsiao presented the PTM[7] based on Chebyshev's inequality as follows [25]:

[7] -Probabilistic Thresholding Method

First step: Discarding the outlier data from abnormal group

In each sample, the abnormal observations (abnormal group) which their MDs are smaller than $\mu_{MDnormal} + 3\sigma_{MDnormal}$ can be eliminated. It is shown by λ ($W > \lambda > 0$). The upper limit for the possibility of a wrong alert is $P(T \le X_{MD}) \le 1 - (W - \lambda)$.

Second step: Calculating the percentage of non-overlapping for normal group

In each sample, thr percentage of normal observations (normal group) in which their MDs are smaller than the smallest MD of abnormal observations (abnormal group) can be calculated. It is shown by W.

Third step: Using Chebyshev's inequality By specifying the values of μ_{MD}, σ_{MD}, W and λ the probabilistic threshold value can be calculated as follows:

$$T = \mu_{MD} + \sigma_{MD} * \sqrt{\frac{1}{1-(W-\lambda)}} \tag{10}$$

Then, the future observations' MDs are compared with the probabilistic threshold value. If they are greater than the probabilistic threshold value, they will be classified into the abnormal group and if they are smaller than the probabilistic threshold value, they will be classified into the normal group. If they are equal to the probabilistic threshold value, they are neither normal nor abnormal (According to the decision-maker, these kinds of observations can be placed into the normal or abnormal group).

3.2 Proposed Model

As stated before, in the Mahalanobis Taguchi system, there are two kinds of misclassification. First, observations those are actually related to the normal group but are categorized into the abnormal group mistakenly and are shown by En_1. Second, observations those are actually related to the abnormal group but are categorized into the normal group mistakenly and are shown by En_2. These misclassifications will reduce the quality of prediction. In this investigation, we will define an upper bound for calculating the En_1 , a lower bound for calculating the En_2, and adding a constraint that will reduce the misclassification and increase the accuracy of prediction.

3.2.1 The Rate of Misclassifications

The rate of misclassification for the normal group is shown by R_N . It is equal to the ratio of En_1 to the total number of observations in the normal group. The rate of misclassification for the abnormal group is shown by R_{AN} . It is equal to the ratio of En_2 to the total number of observations in the abnormal group. The upper bound and lower bound are defined as follows:

$$Lower\,bound = Mmd_N - (ra/r) * Diffmean + Std_N \tag{11}$$

$$Upper\,bound = Mmd_{AN} + (ra/r) * Diffmean - Std_{AN} \tag{12}$$

En_1 and En_2 are calculared as follows:

$$En_1 = \sum_{j=1}^{n_1} \mathrm{MD}_N \quad s.t. \quad MD_N \geq Upper\ bound \tag{13}$$

$$En_2 = \sum_{j=1}^{n_2} \mathrm{MD}_{AN} \quad s.t. \quad MD_{AN} \leq Lower\ bound \tag{14}$$

And the rate of misclassifications is calculated as follows:

$$R_N = \frac{En_1}{r} \tag{15}$$

$$R_{AN} = \frac{En_2}{ra} \tag{16}$$

Decision variables, form a binary p-dimensional vector $X = (x_1, x_2, \dots, x_p)^t$. The x_i Represents the variables of the multivariate system that can take the values 0 or 1. If x_i not selected, it is equal to 0 and if x_i selected, it is equal to 1.

3.2.2 The Mathematical Model

The objective function of the model consists of two parts. The $f_1(X)$ reduces the rate of misclassification and the $f_2(X)$ reduces the number of selected variables.

$$\min f(x) = \alpha\left(W_N \frac{En_1}{r} + W_{AN} \frac{En_2}{ra}\right) + \beta\left(\frac{P_{select}}{P}\right) \tag{17}$$

$$s.t.$$

$$\sum_{i=1}^{p} x_i \leq p \tag{18}$$

$$\sum_{i=1}^{p} x_i = p_{selected} \tag{19}$$

$$MD_{AN} \geq 0.5 + MD_N \tag{20}$$

$$f_1(X) \leq f_1^{max} \tag{21}$$

$$x_i = 0,1$$

The $f_1(X)$ and $f_2(X)$ are positive numbers between 0 and 1. The objective function states that the selected variables must be selected in such a way that the misclassification reduces by the degree of α importance and the number of selected variables reduces by the degree of β importance. Equation (18) states that the sum of the selected variables must be smaller or equal to the total number of variables in the entire data set. When a subset of variables is selected to prepare the measurement scale, the sum of the selected variables is significantly lower than the initial number of variables. Equation (19) states a subset of variables. Equation (20) states that the mean of MDs for the abnormal group must be greater than the mean of MDs for the normal group. Equation (21) states that the total rate of misclassification that is generated by a subset of variablesmust be smaller or equal to when all variables are present in the model.

3.2.3 The Solving Method of the Proposed Model

The proposed model is not classified into the NP-Hard class, because the number of variables is limited and according to the papers that have been presented in this area, there have been no more than 40 variables so far. This model can be solved by MATLAB software (R2012a) and a computer system with Core(TM) i5 - 2410M - CPU 2.30GHz and its time information and computational complexity can be derived. Then, MiniTab16 can gain a fitness function for its prediction.

4. Case Study, Identifying Bad Credit Clients in Banking System

One of the most important services of banks is providing chequebook to their clients. However, it is possible that some clients refuse to pass their cheques and damage to the bank's reputation. Banks can identify these clients and prevent negative consequences. By using the information that has been obtained from a private bank of the Islamic Republic of Iran[8] about the number of bounced cheques, we are trying to evaluate our model and help banks identify and predict the bad credit clients. According to banking experts, ten main factors have an impact on bounced cheques: gender, age, Cheque value, cash date, bounce date, legal status, and number of cheques that have been bounced before, education, job status and payable to. Obviously, these factors are affecting each other. For example, the Cheque value is affected by age, and clients with formal jobs have fewer bounced cheques. The number of bounced cheques and clients' information is in hand for 2017 and 2018. First, we will select the best subset of variables and moderate damage of each bad credit client will be calculated for 2017. Then, the average moderate damage, which has been calculated, will be compared with the real average moderate damages of 2017. Second, we will predict the bad credit clients and their moderate damages for 2018. Then, the average moderate damages, which have been calculated, will be compared with the real average moderate damage of 2018. The mentioned steps will be done by using both, the proposed model and the Taguchi method. Then, their results will be compared. The real average moderate damage is equal to 16168091.59 Rials for 2017 and 17532752 Rials for 2018.

4.1 Solving the Problem by the Proposed Model

There are 162350 bounced cheques in 2017, which have been not passed until 24 March 2017. They may have passed after this time, but, if they could not be passed until 24 March, they are assumed as abnormal observations. Due to a large number of observations, Cochran's test with an error of 0.03 and reliability of 0.95 has been used and 1060 samples were selected randomly. 303 of them (ra=303) have been detected as the abnormal observations (bounced cheques) and placed into the abnormal group, 757 of them (r=757) have been detected as the normal observations (passed cheques) and placed into the normal group. Suppose thatC_1=7 and C_2=3 monetary unit. So, $W_N =0.7$ and $W_{AN} =0.3$. During problem solving, different subsets of variables are considered. For each subset, the Mahalanobis space is formed. As an example, for the subset $(x_1, x_2, x_4, x_5, x_6)$, the MS

[8] 8- Due to banking policies, it is not satisfied with its name in this investigation

is formed based on five variables and 757 normal observations. In order to calculate the abnormal group's MDs, the mean, standard deviation, and correlation matrix of the normal group are used. Finally, the values for the subset are gained as follows:

Std_{AN} =1.7478 Std_{N} =0.5985 Diffmean =1.3272 Mmd_{N} =0.9986

Mmd_{AN} =2.3258 Lower_bound =1.0234 Upper_bound =1.1540

Table1. Defining the variables

Variable		Range
Gender	x_1	1 – 2 (Male – Female)
Age	x_2	20 – 78 years old
Cheque value	x_3	600,000 – 15,470,000,000 Rials
Cash date	x_4	1 – 24 (Table 2)
Bounce date	x_5	1 – 2 (Same day – Next days)
Legal status	x_6	1 – 2 (Court – Non-Court)
Number of Cheques which have been bounced before	x_7	0 – 11
Education	x_8	1 – 6 (Low educated – Diploma – Associate Degree – Bachelor– Master – Ph.D)
Job status	x_9	1 – 3 (Self-employed – Formal – Conditional)
Payable to	x_{10}	1 – 3 (Bearer – Payable to someone – Payable to a company)

Table2. Cash date's ranges

Period	Veriable's value	Period	Veriable's value
March 25 to Apr 5	1	September 25 to October 5	13
April 6 to April 24	2	October 6 to October 24	14
April 25 to May 5	3	October 25 to November 5	15
May 6 to May 24	4	November 6 to November 24	16
May 25 to June 5	5	November 25 to December 5	17
May 6 to June 24	6	December 6 to December 24	18
June 25 to July 5	7	December 25 to January 5	19
July 6 to July 24	8	January 6 to January 24	20
July 25 to August 5	9	January 25 to February 5	21
August 6 to August 24	10	February 6 to February 24	22
August 25 to September 5	11	February 25 to March 5	23
September 6 to September 24	12	March 6 to March 24	24

According to the Equations 13 and 14, the value of En_1 and En_2 are equal to 200 and 64 respectively. Thus, R_N = 0.2642 and R_{AN} =0.2112 and the objective function can be calculated as follows:

$$f_1(x) = W_N R_N + W_{AN} R_{AN} = 0.7*0.2642+0.3*0.2112=0.2483$$
$$F(x) = \alpha * f_1(x) + \beta * f_2(x) = \alpha *0.2483+ \beta *(5/10)$$

α and β can be set between 0 and 1 and their values depend on the decision-maker. If the decision-maker is looking for reducing the misclassification and increase the accuracy, α should be greaterthan β. If the decision-maker is looking for reducing the computational complexity, β should be greaterthan α. This process is repeated for all possible subsets. The subset that has the lowest value of the objective function is known as the best subset. Based on Table 3, the best subset of variables is $\{x_4, x_6, x_7, x_9\}$, which are cash date, legal status, number of cheques which have been bounced before, job status, with $\{\alpha, \beta\}$ = {0.9, 0.1}.

Table3. Candidated subsets with different values of α and β

Value of objective function	Subset	$\{\alpha, \beta\}$
0.2167	$\{x_4, x_6, x_7, x_9\}$	{0.9 , 0.1}
0.2299	$\{x_4, x_6, x_7\}$	{0.8 , 0.2 }
0.2367	$\{x_6, x_7\}$	{0.5 , 0.5}

Now, the average moderate damage will be predicted for 2017. First, the new MS is formed witha selected subset. Second, the abnormal group's MDs are calculated by mean, standard deviation and correlation matrix of the normal group. Next, the probabilistic threshold value is calculated. Based on section *3.1.5*, $\lambda = 0.69$, W=0.0014 and consequently T=1.46 are gained. For better comparison, two optional probabilistic threshold value are considered to be $T_2 = T + \frac{1}{2}\sigma_{MDnormal}$ =1.7584 and $T_3 = T + \sigma_{MDnormal}$ =2.06. The average moderate damage for 2017 and 2018 is shown in table 4 and 5. Moderate damage can be calculated as follows:

$$\text{Moderate Damage} = \frac{\sum_{i=1}^{n} MD_i * Dam_i}{\sum_{i=1}^{n} MD_i} \qquad (22)$$

In which:

n= the number of clients which are examined at a certain probabilistic threshold value Dam_i = the moderate damage of the ith client

Table4. Predicting the average moderate damage for 2017 by proposed model

Average moderate damage (Rials)	Percentage of clients' coverage	Accuracy	Probabilistic threshold value
15,355,765.43	24%	83.7%	1.46
14,556,183.71	21%	73.47%	1.7584
14,825,401.77	18.45%	64.55%	2.06

Table5. Predicting the average moderate damage for 2018 by proposed model

Average moderate damage (Rials)	Probabilistic threshold value
16,979,257	1.46
15,237,482	1.7584
13,972,518	2.06

4.2 Solving the Problem by Taguchi Method

According to Table 6, the Taguchi's orthogonal arrays Table, considers 16 samples for selecting the useful variables among ten variables. As already mentioned, level 1 indicates the presence of that variable in the best subset and level 2 indicates the absence of that variable in the best subset. For each column (variable) the Gain is calculated as Table 7.

The variables $\{x_1, x_4, x_5, x_6, x_8\}$ that are gender, cash date, bounce date, legal status, and education are selected as useful variables and they are placed into the best subset. The results can be observed in Table 8 and 9.

5. Summary and Conclusion

In the proposed model, weight coefficients in relation to the cost are used, but in the Taguchi method, the weight coefficients are not used. Therefore, comparing these two methods may be challenged. However, both methods can be applied to all multivariate systems that MTS can be used for them. In the proposed model, cost-related weights can be ignored. It means that we can consider the same cost-related weights for both parts. By using the proposed model, the four variables have been selected which are cash date, legal status, number of cheques which have been bounced before and job status. By using the Taguchi method, five variables have been selected which are gender, cash date, bounce date, legal status, and education. The proposed model could solve the problem in 0.112 seconds, while, this was more than 7 seconds with the Taguchi method because it has high computational complexity. Meta-heuristic algorithms cannot be used in this case because their accuracy is not desirable. Selecting fewer variables, short solving period, high accuracy, less coding, and computational complexity are the advantages of the proposed model. The proposed model examines all possible subsets, while the Taguchi method examined a limited number of them. In fact, the proposed model discards the subsets which their variables have been repeated into other subsets as same and does not discard any unique subset. The results comparison between the two methods is summarized in Table 10.

Table6. Taguchi's orthogonal arrays Table

Sample	Gender	Age	Cheque value	Cash date	Bounce date	Legal status	cheques which have been bounced	Education	Job status	Payable to	S/N ratio
1	1	1	1	1	1	1	1	1	1	1	-10.18
2	1	1	1	1	1	1	1	2	2	2	-7.83
3	1	1	1	2	2	2	2	1	1	1	-10.9
4	1	1	1	2	2	2	2	2	2	2	-14.48
5	1	2	2	1	1	2	2	1	1	2	-5.94
6	1	2	2	1	1	2	2	2	2	1	-4.93
7	1	2	2	2	2	1	1	1	1	2	-4.75
8	1	2	2	2	2	1	1	2	2	1	-2.47
9	2	1	2	1	2	1	2	1	2	1	-5.59
10	2	1	2	1	2	1	2	2	1	2	-3.37
11	2	1	2	2	1	2	1	1	2	1	-11.73
12	2	1	2	2	1	2	1	2	1	2	-8.52
13	2	1	1	1	2	2	1	1	2	2	-11.19
14	2	1	1	1	2	2	1	2	1	1	-11.67
15	2	1	1	2	1	1	2	1	2	2	-4.79
16	2	1	1	2	1	1	2	2	1	1	-11.14

Table7. The calculation of Gain

Gender	Age	Cheque value	Cash date	Bounce date	Legal status	Number of cheques which have been bounced before	Education	Job status	Payable to	The average of level
-3.84	-4.54	-5.14	-3.79	-4.07	-3.13	-4.27	-4.07	-4.15	-4.29	1
-4.25	-3.55	-2.96	-4.30	-4.13	-4.96	-3.82	-4.12	-3.94	-3.80	2
0.41	-0.98	-2.18	0.5	1.83	0.06	1.83	-0.45	-0.22	-0.48	Gain

Table8. Predicting the average moderate damage for 2017 by Taguchi method

Average moderate damage (Rials)	Percentage of clients' coverage	Accuracy	Probabilistic threshold value
14,960,950	14.6%	51%	1.36
12,275,536.1	11%	38%	1.61
11,816,164.1	8%	27%	1.87

Table9. Predicting the average moderate damage for 2018 by Taguchi method

Average moderate damage (Rials)	Probabilistic threshold value
15,138,026	1.36
12,831,294	1.61
11,984,694	1.87

Table10. Results comparison of the proposed model with the Taguchi method

Solving by	Selected variables	Probabilistic threshold value	Accuracy	Prediction error for 2018	The total squared error devided by 10,000,000
The proposed model	cash date, legal status, number of cheques which have been bounced before, job status	1.46	84%	16,979,257	1,824,988.7
		1.76	73%	15,237,482	
		2.06	65%	13,972,518	
The Taguchi method	gender, cash date, bounce date, education	1.36	51%	15,138,026	5,861,936.7
		1.61	38%	12,831,294	
		1.87	27%	11,984,694	

For an accurate prediction and detection of normality and abnormality of future observations, the concept of reliabilitycan is used. It will lead to a precise threshold value and appropriate to various systems. Moreover, the MTS can be used for optimizing models. For example, use it to improve the off-line parameters in the Meta-heuristic algorithms such as the GA.

6. Refrences

[1] Guyon, I. and Elisseeff, A. 2003. An Introduction to Variable and Feature Selection. Journal of machine learning research. 3: 1157-1182.

[2] Khan, S. 2008. Alpaydin Ethem. Introduction to Machine Learning (Adaptive Computation and Machine Learning Series). The MIT Press, 2004. ISBN: 0 262 01211 1 Price£ 32.95/$50.00 (hardcover). xxx+ 415 pages. Natural Language Engineering. 14(1): 133-137.

[3] Ng, A.Y. 2004. Feature Selection, L1 vs. L2 regularization, and rotational invariance. In Proceedings of the twenty-first international conference on Machine learning, 78.

[4] Taguchi, G., Chowdhury, S. and Wu, Y. 2005. Taguchi's Quality Engineering Handbook (Vol. 1736). Hoboken, NJ: John Wiley & Sons.

[5] Lee, Y. C. and Teng, H. L. 2009. Predicting the Financial Crisis by Mahalanobis–Taguchi system–Examples of Taiwan’s Electronic Sector. Expert Systems with Applications. 36(4): 7469-7478.

[6] Das, P. and Datta, S. 2007. Exploring the Effects of Chemical Composition in Hot Rolled Steel Product using Mahalanobis Distance Scale under Mahalanobis–Taguchi System. Computational Materials Science. 38(4): 671-677.

[7] Rai, B.K., Chinnam, R. B. and Singh, N. 2008. Prediction of Drill-bit Breakage from Degradation Signals using Mahalanobis-Taguchi System Analysis. International Journal of Industrial and Systems Engineering. 3(2): 134-148.

[8] Yang, T. and Cheng, Y.T. 2010. The Use of Mahalanobis–Taguchi System to Improve Flip-chip Bumping Height Inspection Efficiency. Microelectronics Reliability. 50(3): 407-414.

[9] Abbasi, S. E., Aaghaie, A. and Fazlali, M. 2011. Applying Mahalanobis–Tagouchi System in Detection of High Risk Customers–A Case-based Study in an Insurance Company. 45(Special Issue): 1-12.

[10] Jin, X. and Chow, T. W. 2013. Anomaly Detection of Cooling Fan and Fault Classification of Induction Motor Using Mahalanobis–Taguchi System. Expert Systems with Applications. 40(15): 5787-5795.

[11] Pal, A. and Maiti, J. 2010. Development of a Hybrid Methodology for Dimensionality Reduction in Mahalanobis–Taguchi System Using Mahalanobis Distance and Binary Particle Swarm Optimization. Expert Systems with Applications. 37(2): 1286-1293.

[12] ReséNdiz, E. and Rull-Flores, C. A. 2013. Mahalanobis–Taguchi System Applied to Variable Selection in Automotive Pedals Components Using Gompertz Binary Particle Swarm Optimization. Expert Systems with Applications. 40(7): 2361-2365.

[13] Barahimi, A.H. and Aghaie, A. 2017. Modeling of Integer Programming for Selection of Useful Variables in a Multivariables System Using the Mahalanobis-Taguchi System، Case study: DAMAGE, 12th International Industrial Engineering Conference2017, Kharazmi University, Tehran, IRAN). (In Persian)

[14] Holcomb, S. 2017. Mahalanobis Taguchi System (MTS) for Pattern Recognition, Prediction, and Optimization. International Conference of Modeling and Simuation. 34: 1-8.

[15] Li, C.B., Yuan, J.H. and Gao, P. 2017. Risk Decision-making Based on Mahalanobis-Taguchi System and Grey Cumulative Prospect Theory for Enterprise Information Investment. Intelligent Decision Technologies. 10(1): 49-58.

[16] Okafor, E.C., Ihueze, C.C. and Nwigbo, S.C. 2013. Optimization of Hardness Strengths Response of Plantain Fibres Reinforced Polyester Matrix Composites (pfrp) Applying Taguchi Robust Resign. International Journal of Science & Emerging Technologies. 5(1): 217-227.

[17] Singh, R., Rizvi, S. A. and Tewari, S. P. 2018. Effect of Friction Stir Welding on the Tensile Properties of AA6063 Under Different Conditions. International Journal of Engineering Transactions A: Basics. 30(4): 597-603.

[18] Taguchi, S. 2000. Mahalanobis-Taguchi System, ASI Taguchi Symposium.

[19] Aman, H., Mochiduki, N. and Yamada, H. 2006. A Model for Detecting Cost-Prone Classes based on Mahalanobis-Taguchi Method. IEICE Transactions on Information and Systems. 89(4): 1347-1358.

[20] Foley, D. 1972. Considerations of Sample and Feature Size. IEEE Transactions on Information Theory. 18(5): 618-626.

[21] Leese, M. N. and Main, P. L. 1994. The Efficient Computation of Unbiased Mahalanobis Distances and Their Interpretation in Archaeometry. Archaeometry. 36(2): 307-316.

[22] Mohan, D., Saygin, C. and Sarangapani, J. 2008. Real-time Detection of Grip Length Deviation During Pull-type Fastening: a Mahalanobis–Taguchi System (MTS)-based Approach. The International Journal of Advanced Manufacturing Technology. 39(9-10): 995-1008.

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

[24] Stoumbos, Z. G., Koudelik, R., Tsui, K. L., Kim, S. B., Woodall, W. H. and Carvounis, C. P. 2004. A Review and Analysis of the Mahalanobis-Taguchi System. Quality Control and Applied Statistics. 49(2): 195-198.

[25] Su, C. T. and Hsiao, Y. H. 2007. An Evaluation of the Robustness of MTS for Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering. 19(10): 1321-1332.