Expert Systems with Applications 41 (2014) 8003–8015

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Enhancement of Mahalanobis–Taguchi System via Rough Sets based Feature Selection

Ashif Sikandar Iquebal [a], Avishek Pal [b,*], Darek Ceglarek [b], Manoj Kumar Tiwari [a]

[a] Department of Industrial Engineering & Management, IIT Kharagpur, Kharagpur, 721302, India
[b] WMG, University of Warwick, Coventry, CV4 7AL, United Kingdom

ARTICLE INFO

Article history:
Available online 18 June 2014

Keywords:
Data mining
Mahalanobis Taguchi System
Feature Selection
Orthogonal Arrays
Rough Sets
Over-fitting
Regularization
Conditional probability
IF–THEN rules

ABSTRACT

The current research presents a methodology for classification based on Mahalanobis Distance (MD) and Association Mining using Rough Sets Theory (RST). MD has been used in Mahalanobis Taguchi System (MTS) to develop classification scheme for systems having dichotomous states or categories. In MTS, selection of important features or variables to improve classification accuracy is done using Signal-to-Noise (S/N) ratios and Orthogonal Arrays (OAs). OAs has been reviewed for limitations in handling large number of variables. Secondly, penalty for over-fitting or *regularization* is not included in the feature selection process for the MTS classifier. Besides, there is scope to enhance the utility of MTS to a classification-cum-causality analysis method by adding comprehensive information about the underlying process which generated the data. This paper proposes to select variables based on maximization of *degree-of-dependency* between Subset of System Variables (**SSV**s) and system classes or categories (**R**). *Degree-of-dependency*, which reflects goodness-of-model and hence goodness of the **SSV**, is measured by conditional probability of system states on subset of variables. Moreover, a suitable *regularization* factor equivalent to L_0 norm is introduced in an optimization problem which jointly maximizes goodness-of-model and effect of *regularization*. Dependency between **SSV**s and **R** is modeled via the equivalent sets of Rough Set Theory. Two new variants of MTS classifier are developed and their performance in terms of accuracy of classification is evaluated on test datasets from five case studies. The proposed variants of MTS are observed to be performing better than existing MTS methods and other classification techniques found in literature.

1. Introduction

Classification is a predictive modeling technique which is used to predict membership of objects to system states or categories (**R**). Category membership depends on properties of the objects which are described by attributes or variables (**X**) also known as features in machine learning parlance. Classification techniques develop analytical models to represent the dependency of categories on variables. Models are built from training dataset where category membership and values of system variables are known for given objects. Many classification techniques apply a **Predictor Function**, $f(x)$ to find class labels which represent system states. Objects (x) are treated as input to the predictor function which gives the class labels as output (y). When the output y is continuous then additional steps are required to segment the continuous scale into intervals equivalent to discrete class labels that can be assigned to the objects. In a two-state system, class labels $\mathbf{CL} = \{CL_1, CL_2 \ldots CL_q\}$ and states $\mathbf{R} = \{R_1, R_2\}$ have *one-to-one* mappings which are represented as $CL_i \rightarrow R_i$, where $i = 1, 2$. Misclassification happens when the **Predictor Function** classifies an object $x \in CL_i$ while in reality $x \in R_j$ (therefore $i \neq j$). Often the predicted class CL_i is used to draw inference about state or behavior of the object, such as detecting illness from medical images (Antonie, Zaiane, & Coman, 2001; Lavrač, 1999) or predicting fraudulent activity from a loan applicant's records (Chan, Fan, Prodromidis, & Stolfo, 1999; Phua, Lee, Smith, & Gayler, 2010). Hence, in case of decision making, misclassification is costly and therefore should be eliminated or reduced. Therefore a subset of variables which maximizes predictive accuracy is sought. Percentage or fraction of correct classification is a measure of goodness-of-model. However, selecting subset of variables solely based on improving classification accuracy leads to over-fitting whereby the model predicts better on training data but fails to generalize for data other than training (Alpaydin, 2004; Dietterich, 1995). Therefore, appropriate steps should be taken during feature selection to avoid over-fitting on training data set. Additionally, important variables, which are used to build the final model, provide valuable understanding of the underlying

[*] Corresponding author. Tel.: +44 (0) 246 761 50759.
E-mail address: avishek.pal@warwick.ac.uk (A. Pal).

http://dx.doi.org/10.1016/j.eswa.2014.06.019
0957-4174/© 2014 Elsevier Ltd. All rights reserved.

process or behavior which generated the data (Guyon & Elisseeff, 2003). Knowledge of the process or behavior helps in targeted decision making such as promoting purchase (Ahn, Ahn, Byun, & Oh, 2011) and enabling cross-selling (Ahn, Ahn, Oh, & Kim, 2011) of telecom products to specific customers based on their preferences. Therefore, **Feature Selection** is important to identify variables which improve the accuracy of the predictor function, avoid over-fitting of the model on training data and provide a comprehensive description of the causal relationship between variables and predicted classes. Besides a predictor function on fewer variables is computationally more efficient.

Mahalanobis Taguchi System (MTS) (Taguchi & Jugulum, 2002; Woodall et al., 2003) is a classification methodology which uses the notions of **Predictor Function** and **Feature Selection**. MTS is used for classification of dichotomous systems having two class labels **CL** = {CL_1, CL_2} which represent a two-state system for example **R** = {normal, abnormal}, **R** = {accepted, rejected} etc. MTS has been applied in many fields such as health care for medical diagnosis, manufacturing for quality inspection and fault detection, finance for bankruptcy prediction and several other areas. MTS uses Mahalanobis Distance (MD) to construct Mahalanobis Space (Taguchi & Jugulum, 2002) based on objects belonging to a particular category for example '*normal*'. Average of MD values of '*normal*' objects is close to 1 and any object whose MD is significantly more than a threshold Mahalanobis Distance (MD_T) is classified to belong to '*abnormal*' category. The threshold MD is optimally set to a value which minimizes percentage misclassification by the Mahalanobis Space determined by it. Besides, MD has a great advantage over Euclidean Distance or Manhattan Distance (Aggarwal, Hinneburg, & Keim, 2001) as MD takes into account the correlation between variables which effects dependency of categories (**R**) on system variables (**X**). For feature selection, MTS uses Orthogonal Arrays (OAs) to identify Subset of System Variables, $\mathbf{SSV} \subseteq \mathbf{X}$ which maximizes goodness-of-model measured by Signal-to-Noise (S/N) ratios (Taguchi, Chowdhury, & Wu, 2005; Taguchi & Jugulum, 2002). However several researches have reviewed the limitations of OAs and S/N ratios for variable selection (Abraham & Variyath, 2003). Besides OA's as suggested in Taguchi's fractional factorial experiment design are applicable to limited number of variables and levels (Orthogonal Arrays, 2004). Moreover, for large number of variables, OAs is not able to evaluate all possible subsets of features. To address the limited search capability of OAs, Pal and Maiti (2010) addressed the feature selection problem as binary integer programming minimizing total weighted misclassification (TWM) via Binary Particle Swarm Optimization (BPSO). There are numerous other works applying the MTS classifier and selecting important variables either by OAs or optimization by solely improving goodness-of-model. However there is lack of research on addressing the issue of over-fitting during feature selection for the MTS classifier. This paper introduces a penalty on over-fitting and develops a feature selection methodology which not only maximizes model fitness but also tackles over-fitting due to inclusion of too many features. Moreover, a new indicator of goodness-of-model, which is measured in terms of '*degree of dependency*' between system states and subset of variable, is explored. The proposed methodology is applied on five datasets and prediction accuracy is checked on test datasets, which are not used to train the model. In several cases, the new feature selection method gives MTS classifier, which performs better than the ones obtained from feature selection methods such as maximizing S/N ratio by OAs or minimizing TWM via BPSO. Additionally, the present methodology extracts causal relationship between critical variables and system states in terms of IF-THEN association rules, which leads to better understanding of the underlying process and behavior. This enhances utility of MTS to classification tool cum causality analysis tool based on few critical attributes.

The rest of the paper is organized as follows. Section 2 describes related work on feature selection for the MTS classifier. This is followed by a detailed description of the proposed methodology in Section 3. Next, Section 4 presents the case studies and results. Performance of the current methodology is compared with other methods based on accuracy of classification. Finally, the paper concludes with a discussion of the overall research contribution and remarks on future work.

2. Literature Review

Feature Selection has been an area of considerable interest in machine learning with increasing availability of problems and datasets having large number of variables. Guyon & Elisseeff, 2003 explains that the importance of variables can be evaluated in terms of (i) correlation with the output (y); and (ii) information entropy of output conditional to input variables (x). Subset of variables $\mathbf{SSV} \subseteq \mathbf{X}$ ($p = |X| \& q = |\{SSV\}|$) can be selected based on rankings from these measures of feature importance or from direct optimization of competing objectives: (i) maximization of goodness-of-model evaluated by indicators such as coefficient of determination, R^2 or log-likelihood, which is calculated from residual sum squares, *RSS*; and (ii) Minimization of number of variables for example through L_0 norm ($2q$) in Akaike Information Criterion (Bozdogan, 1987), or by L_1 norm ($\sum_{i=1}^{q}\|\beta_i\|^1$) for lasso regression (Tibshirani, 1996) or through L_2 norm ($\sum_{i=1}^{q}\|\beta_i\|^2$) in case of ridge regression (Myers, 1990), where β_i is the regression coefficient or weight of the ith variable in **SSV**. In general kth norm is $L_k = \sum_{i=1}^{q}\|\beta_i\|^k$. Research has shown that maximizing goodness-of-model alone improves prediction accuracy on training dataset but leads to over-fitting and failure of the model to generalize for data outside training (Alpaydin, 2004; Dietterich, 1995). Therefore minimizing number of variables is important to penalize over-fitting and improve prediction accuracy for out-of-training data (Ng, 2004; Schölkopf & Smola, 2001). The penalizing effect of the second objective function to prevent over-fitting is known as *regularization*.

For Mahalanobis–Taguchi System (MTS), feature selection has been done only by improving goodness of the MTS classifier. For example, Taguchi et al., 2005 maximizes Signal-to-Noise (S/N) ratio to select variables. Subsets of variables are generated using Taguchi's Orthogonal Arrays (OAs) and S/N ratio of each such subset is evaluated. Several researches have used S/N ratio and OAs for feature selection to improve predictive accuracy of the MTS classifier in different applications. Lee and Teng (2009) use large number of financial ratios as variables to build MTS classifier that predicts bankruptcy of companies. Critical features are selected through maximization of S/N ratio. MTS has also been used in quality inspection and fault detection in manufacturing. For example, Das and Datta (2007) apply MTS classier to predict quality of hot-rolled steel plates based on chemical composition, whereby elements significant for quality of steel plate are selected by maximizing S/N ratios. Rai, Chinnam, and Singh (2008) develops a MTS classifier to detect breakage of metal drilling tool. S/N ratio and OAs determine important predictors from vibration signal parameters. Yang and Cheng (2010) uses S/N ratio and OAs to identify critical inspection parameters for quality check of flip-chips and determines the quality of the fabrication by MTS classifier. Jin and Chow (2013) develop Mahalanobis Distance (MD) based classification to indicate health of cooling fan and induction motor. Important parameters, which are extracted from vibration signals based on S/N ratio, are used to build MTS classifier that detects fault in induction motor. However, S/N ratio is solely based on *Mahalanobis Distance* (MD) and does not directly consider predictive performance of the model. Besides, OAs cannot generate of all possible subsets of variables. To address these limitations, Pal and Maiti (2010) suggest minimization of total weighted misclassification (TWM) as goodness-of-model and formulates the variable selection process as a binary integer program, which is solved by Particle Swarm Optimization (PSO). Case study on predicting quality of casting from foundry parameters shows that the new feature selection strategy gives better classification accuracy than important variables found by S/N ratio and OAs. Reséndiz and Rull-Flores (2013) apply MTS classifier to predict the quality of automotive pedals made by injection moulding process. Variables to improve prediction accuracy are selected by Gompertz binary particle swarm optimization which reduces TWM of twofold quality markers. Reséndiz, Moncayo-Martínez, and Solís (2013) minimize TWM via binary ant colony optimization to select variables for MTS classifier, which predicts assembly quality of truck body from dimensional parameters. Though a lot of research has been conducted on improving prediction accuracy of MTS classifier by optimizing goodness of model, little has been done to incorporate measures to penalize over-fitting in the feature selection process. This research addresses the lack of *regularization* in feature selection for MTS. The proposed feature selection process introduces an objective function which improves model accuracy and well as avoids over-fitting by adding a *regularization* factor which is equivalent to penalty via L_0 norm. The joint objective function is optimized by Genetic Algorithm to determine the minimal subset of features. Besides, goodness-of-model in the new feature selection method is evaluated by *degree of dependency* or conditional probability of predicted classes on subset of variables. Overall, the developed method includes a new measure of model fit and a penalty on over-fitting. Case studies show that the proposed feature selection renders higher predictive accuracy to the MTS classifier than feature selection methods found in literature.

Another goal of feature selection is to provide knowledge about the underlying process that generated the data. Past research has focused on identifying important variables which improve accuracy of the MTS classifier. However there is lack of an approach to provide a better explanation about how the important variables influence predicted classes (CL) and corresponding system states (R). This research enhances the utility of the MTS classifier by developing a feature selection method which not only identifies critical variables that improve classification accuracy and minimize over-fitting but also extracts IF-THEN causal rules (also known as association rules), which link attribute–value combinations with system states. These rules provide more comprehensive description of the underlying behavioral patterns and effects of the selected variables on the system states. Association rules have received considerable interest as a tool to represent cause-effect relationships. Generating association rules involve identifying attribute–value combinations that show strong correlation, conditional probability or conditional information entropy with responses or predicted classes (Agrawal, Imieliński, & Swami, 1993). Association rules have been applied to analyze customer buying patterns (Berry & Linoff, 1997), track changes in customer purchases (Song & Kim, 2001), recommend products for on-line purchases (Changchien & Lu, 2001), predict presence of illness (Gamberger, Lavrac, & Jovanoski, 1999), map cause-effect relations among demographic factors from census data (Malerba, Esposito, Lisi, & Appice, 2003). In this paper, the association or dependency between variables (**X**) and states (**R**) are modeled by generalized Rough Set Theory (Pawlak, 1982). Identification of important predictor variables through Rough Sets have been done for several application such as medical diagnosis (cardiac illness - Komorowski & Ohrn, 1999, diabetis – Nakayama, Hattori & Ishii, 1999 and breast cancer - Hassanien & Ali, 2004), text categorization (Li, Shiu, Pal & Liu, 2006), web-page classification (An, Huang, Huang & Cercone, 2005), faulty manufacturing parameter detection for warranty failure root cause analysis (Mannar, Ceglarek, Niu & Abifaraj, 2006), generation of diagnostic decision rules for repair of warranty failures (Pal & Ceglarek, 2013), etc. Thangavel and Pethalakshmi (2009) present a review of application of Rough Set theory for feature selection.

In summary, the current research goes beyond the state-of-the-art in MTS by making contributions in 3 folds: (i) introduction of a new measure of goodness of the MTS classifier in terms of conditional probability of predicted classes on subset of variables; (ii) addition of *regularization* or penalty for over-fitting in the feature selection process; and (iii) identification of cause-effect relationships between selected variables and predicted classes in terms of IF–THEN rules, which enhance the utility of MTS from a classification tool to a predictive modeling cum causality analysis method for systems with discrete states. Table 1 presents a synopsis of comparison between current approaches and the proposed methodology of feature selection for the MTS classifier.

Table 2 summarizes the classification techniques with which prediction performance of the proposed method has been compared in the five case studies. The summary in Table 2 highlights the **Prediction Function** and **Feature Selection** scheme used by each these methods.

3. Proposed methodology

This section describes the proposed methodology in a holistic manner whereby the new feature selection process is embedded as an integral part in the overall framework of Mahalanobis–Taguchi System (MTS). This facilitates application of a single

Table 1
Comparison of state-of-art in feature selection for MTS and proposed methodology.

Topic		Methodology	
		State-of-art	Proposed
Objective function for feature selection	Goodness-of-model	- Maximization of Signal-to-Noise ratio [9–13 & 23–27] - Minimization of total weighted misclassification [15, 28 & 29]	- Maximization of causal dependency of predicted classes on variables - Causal dependency measured by conditional probability of predicted classes on variables
	Regularization	×	- Maximization of fraction of variables screened, $\frac{p-q}{p}$, where p is total number of variables and q is number of variables in subset - Equivalent to L_0 norm penalty for over-fitting
	Optimization technique	- Orthogonal Arrays [9–13 & 23–27] Binary particle swarm optimization [15 & 28] - Binary ant colony optimization [29]	Genetic algorithm
	Process knowledge	×	IF–THEN causal rules linking selected variables and predicted classes

Table 2
Related work on classification using predictor function & feature selection. Bold values highlights the proposed methodology.

Methods	Classification steps		
	Predictor function	Feature selection	
		Criterion for variable selection	Technique of variable selection
MTS-PSO (Pal & Maiti, 2010)	Mahalanobis distance (MD)	Maximum Signal-to-Noise (S/N) ratio	Orthogonal Arrays (OAs)
MTS-Orthogonal Arrays (OAs) (Taguchi & Jugulum, 2002)		Total weight cost of misclassification (TWCM)	Binary integer programming using particle swarm optimization
Proposed Methodology **Mahalanobis Distance & rough sets based association mining for classification**		**Maximum strength of association between Subset of System Variables & categories**	**Genetic algorithm**
3-NN Stand Manhattan (Duch, Adamczak, Grabczewski, & Zal, 1998)	Total coverage of set of logical rules	Generation of set of rules that predicts category memberships	Constrained multi-layer perceptron neural network
Naïve Bayesian classifier (Al-Aidaroos, Bakar, & Othman, 2010)	Classification based on Bayes theorem assuming strong independence	Rough Set Theory was used to remove the redundant variables from the system	
Bayesian pair-wise classifier (Duch et al., 1998)	Posterior probabilities	Variables selection is not done	
Feature space mapping (Duch et al., 1998)	Total coverage of a set of logical rules	Generation of a set of rules that decides the decision class	
Fisher linear discriminant analysis (Durrant & Kabán, 2010)	Separating samples using line projection	Variables selection is not done	
C4.5 (Decision Tree) (Bredensteiner & Bennett, 1998)	Decision Trees	Formulation of parametric bilinear sub problems within given misclassification error tolerance	Frank–Wolfe method to solve the sub problems
Ant colony optimization (Parpinelli, Lopes, & Freitas, 2001)	Discovering classification rules inspired by the behavior of real ant colony	Variables selection is not done	
Rule induction through approximate classification (RIAC) (Hamilton, Shan, & Cercone, 1996)	Generate individual rules with certainty factor based on RIAC and then combine these rules	Coverage of the decision class using a minimal set	Degree of dependency
Neural network (Setiono & Liu, 1997)	Neural network	Removal of attributes based on the accuracy rate of the network	Three layered feed forward neural network

enhanced MTS classifier which includes a new feature selection method. The architecture of the improved MTS classifier is based on the following generic notions:

(i) *Development of Predictor function* – To develop the predictor function, objects belonging to a particular category (R_i) are used from the training dataset. For example the predictor function is developed using objects belonging to R_i = normal, R_i = accepted or $R_i = 0$ categories. The predictor function $f(x)$ suggested in this paper calculates Mahalanobis Distance (MD) (Taguchi & Jugulum, 2002) for a given object, x and is denoted as MD(x).

(ii) *Segmentation of continuous scale into class labels* – MD is a continuous variable. Therefore optimal cut-off or threshold $\mathrm{MD_T}$, is selected to partition the continuous scale of MD values such that $\mathrm{MD(x)} \leqslant \mathrm{MD_T} \rightarrow \mathrm{CL_1}$ and $\mathrm{MD(x)} > \mathrm{MD}_T \rightarrow \mathrm{CL_2}$. $\mathrm{MD_T}$ is selected such that class labels obtained from it gives minimum percentage misclassification of the training dataset.

(iii) *Selection of optimal Subset of System Variables (SSV_{opt})* – The methodology finds $\mathbf{SSV_{opt}}$ based on strength of association between **SSV**s and system categories $\mathbf{R} = \{R_1, R_2\}$. Strength of association (Bell, Guan, & Liu, 2005) is used to measure *'goodness'* of IF–THEN rules in predicting categories. Current methodology suggests an index to measure strength of association between **SSV**s and system categories **R**. This index is termed as *degree of dependency* $\delta_{\mathrm{SSV}}(\mathbf{R})$, which is based on the conditional probability of the categories in the system variables. However, maximizing only goodness of model δ_{SSV} (**R**) through variable selection leads to over-fitting on training data. Therefore, a penalty factor on the number of variables is added to the objective function. To solve feature selection as a joint maximization problem, the penalty factor is actually represented as the fraction of variables screened or eliminated to build the MTS classifier.

(iv) *Rebuilding of predictor function and output class labels* – The predictor function is built using $\mathbf{SSV_{opt}}$ and new class labels are obtained. Percentage misclassification based on the new class labels is re-evaluated.

The steps which implement the aforementioned notions are shown in the flowchart of Fig. 1. This is followed by detailed description of the steps.

The steps of the methodology are detailed as follows.

3.1. Step I: Calculate Mahalanobis Distance (MD) from standard deviations & correlation matrix of 'normal' sample

In multi-dimensional classification, the correlation between the system variables plays an important role in determining the output class labels. This paper applies Mahalanobis Distance (MD) to capture the effect of correlation between variables on class labels. For an object x_i where $i = 1, 2 \ldots n$, in a multivariate system with 'p' variables, the MD is obtained as

$$\mathrm{MD}(x_i) = \frac{1}{n} z_i^T S^{-1} z_i \tag{1}$$

where z_i is normalized vector for object x_i and $z_i = [z_{i1}, z_{i2} \ldots z_{ip}]^T$ such that

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \tag{2}$$

$$\bar{x}_j = \frac{\sum_{i=1}^{n} x_{ij}}{n} \tag{3}$$

S^{-1} is inverse of the correlation matrix of the *'normal'* sample. $\bar{x}_j$ and s_j are the average and standard deviation of the *'normal'* sample where $j = 1, 2 \ldots p$.

MDs calculated using standard deviations and correlation matrix of *'normal'* objects gives a continuous measurement scale

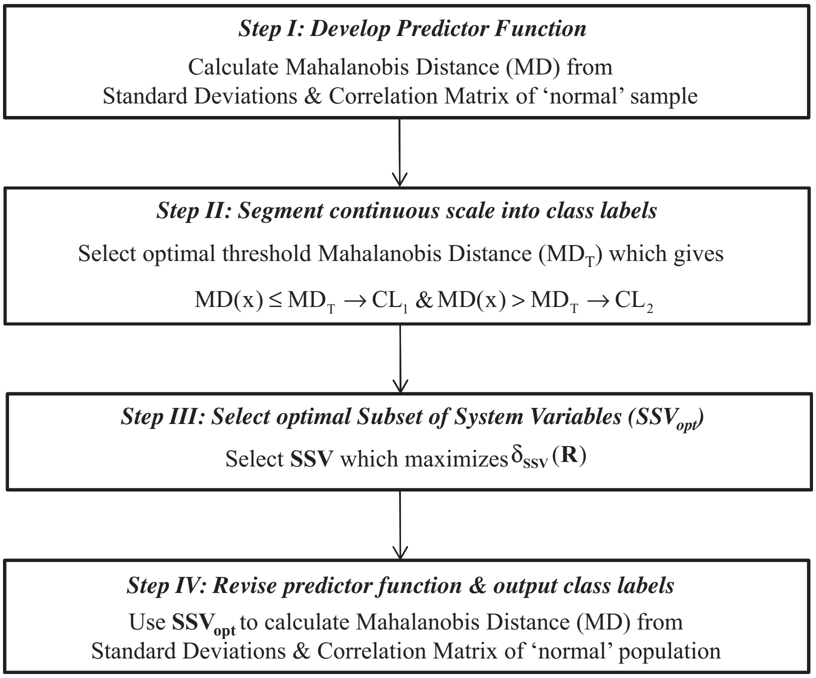
Step I: Develop Predictor Function
Calculate Mahalanobis Distance (MD) from
Standard Deviations & Correlation Matrix of 'normal' sample
Step II: Segment continuous scale into class labels
Select optimal threshold Mahalanobis Distance (MD_T) which gives
$MD(x) \le MD_T \rightarrow CL_1$ & $MD(x) > MD_T \rightarrow CL_2$
Step III: Select optimal Subset of System Variables (SSV_{opt})
Select **SSV** which maximizes $\delta_{SSV}(\mathbf{R})$
Step IV: Revise predictor function & output class labels
Use $\mathbf{SSV_{opt}}$ to calculate Mahalanobis Distance (MD) from
Standard Deviations & Correlation Matrix of 'normal' population

Fig. 1. Steps of proposed classification methodology.

termed as Mahalanobis Space (MS). On this scale, the MDs of *'abnormal'* objects will be significantly larger than those of normal. Whereas the average of *'normal'* MD is close to 1 (Sun, Nguyen, Vu, & Bisland, 2006). However there will be certain overlap which will result in misclassification.

Additionally, the two system categories are denoted as R_1 = normal and R_2 = abnormal. The number of *'normal'* and *'abnormal'* objects in the training dataset is indicated as 'n' and 'n'' respectively.

3.2. Step II: Select optimal threshold Mahalanobis Distance (MD_T) to generate class labels

For an object x, if $MD(x) \leqslant MD_T$, then $x \in CL_1$ else $x \in CL_2$. Misclassification occurs if actually $x \in R_1$ as found in training dataset but $MD(x) > MD_T$ from predictor function. Similarly misclassification also occurs if $x \in R_2$ as found in training dataset though $MD(x) \leqslant MD_T$ from predictor function. Total percentage misclassification (p) is expressed as a function of MD_T by

$$p(MD_T) = \frac{|\{x|x \in R_1 \wedge MD(x) > MD_T\}| + |\{x|x \in R_2 \wedge MD(x) \leqslant MD_T\}|}{n + n'} \times 100 \tag{4}$$

where $\wedge$ is logical AND combining two criteria. MD_T is selected such that based on it minimum percentage misclassification is obtained. To determine MD_T, let us first define MD_{lower} and MD_{upper} as follows.

$$MD_{lower} = \frac{MD_{max} - MD_{min}}{2} \tag{5}$$

$$MD_{upper} = \frac{MD_{max} + MD_{min}}{2} \tag{6}$$

where MD_{max} and MD_{min} are minimum and maximum $MD(x)$ found in training dataset. Boundary Region (BR) is defined as the set of objects whose MD values are greater than or equal to MD_{lower} and less than or equal to MD_{upper}. BR can be obtained by

$$BR = \{x|MD_{lower} \leqslant MD(x) \leqslant MD_{upper}\} \tag{7}$$

MD_T is obtained from $x \in BR$ such that it gives minimum misclassification as calculated by Eq. (4).

3.3. Step III: Select optimal Subset of System Variables (SSV_{opt})

The steps involved in identifying $\mathbf{SSV_{opt}}$ are listed as follows:

A. *Generation of Selection Information System (SIS)* – This step uses the training dataset to generate information system for selection of optimal Subset of System Variables.
B. *Discretization using Equal Frequency Binning & Khiops method* – System variables could be continuous or categorical. This step discretizes continuous variables by using by using Equal Frequency Binning (Slowinski, 1992) and Khiops Discretization method (Kotsiantis & Kanellopoulos, 2006). Results related to both the methods are presented in the comparison section. Variable selection is done based on discretized class intervals obtained from both methods and results of classification are obtained for each case.
C. *Determinations of equivalent classes for SSVs* – For a given **SSV**, equivalent classes are sets of objects having same values for the variables in **SSV**. The values of the system variables are based on the discretized intervals obtained from discretization.
D. *Calculate degree of dependency* $\delta_{SSV}(\mathbf{R})$ – Degree of dependency measures the ability of a **SSV** to differentiate between the two system categories R_1 and R_2. The degree of dependency is calculated based on the concept of membership of objects to equivalent classes.
E. *Identification of optimal SSV (SSV_{opt})* – For a multivariate system of p variables there are $2^n - 1$ candidate solutions for $\mathbf{SSV_{opt}}$. Genetic Algorithm is used to generate candidate solutions randomly and identify $\mathbf{SSV_{opt}}$ based on maximum $\boldsymbol{\delta}_{\mathbf{SSV}}(\mathbf{R})$.

The steps of the variable selection are listed as flowchart in Fig. 2.

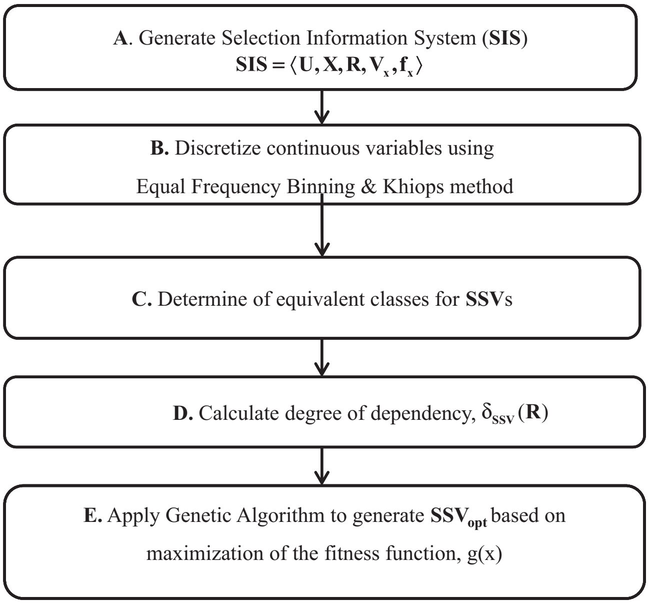
A. Generate Selection Information System (SIS)
SIS = ⟨U, X, R, V_x, f_x⟩
B. Discretize continuous variables using
Equal Frequency Binning & Khiops method
C. Determine of equivalent classes for SSVs
D. Calculate degree of dependency, δ_SSV(R)
E. Apply Genetic Algorithm to generate SSV_opt based on
maximization of the fitness function, g(x)

Fig. 2. Rough sets based feature selection for Mahalanobis–Taguchi System.

The steps involved in selection optimal Subset of System Variables are detailed as follows:

3.3.1. Discretize continuous variables using Equal Frequency Binning & Khiops method

(1) *Equal Frequency Binning algorithm*: The range of values of $x \in X$ is divided into $t-1$ cuts such that the number of objects that fall into each of the t intervals are approximately equal. The advantage of this method that it is simple to use.

(2) *Khiops Discretization method*: Khiops discretization technique (Boulle, 2004; Liu, Hussain, Tan, & Dash, 2002) is constructed based on the optimization of Chi-square value. It initially starts with single valued intervals and then merges them depending upon the Chi-square criterion applied to the training data set. Based on the significance level, the stopping criteria for the merging is decided. The algorithm tends to minimize the decrease in the Chi-square value between the discretized variable and the system categories whenever a merging of intervals is done. During merging two possible cases arise: (i) merging in which the intervals do not have minimum number of elements (ii) merging in which the intervals have minimum number of elements. By default, first type of merging is done.

The algorithm iterates till all the intervals have minimum number of elements. Once all intervals have minimum number of elements, further merging is done till decrease in the Chi-square value is less than a threshold. If the variable and the system category are independent, then the algorithm tends to give a single interval. If probability of converging towards a single interval is p, then the maximum value of decrease in the chi square value below which a merge is accepted is given by MaxDeltaChi2 as

$$P(\mathit{DeltaChi}2 \leqslant \mathit{MaxDeltaChi}2)^N \geqslant p \tag{8}$$

$$P(\mathit{Chi}2_{c-1} \leqslant \mathit{MaxDeltaChi}2) \geqslant p^{1/N} \tag{9}$$

$$\mathit{MaxDeltaChi}2 = \mathit{In}\,v\mathit{Chi}2_{c-1}(\mathit{prob} \geqslant p^{1/N}) \tag{10}$$

3.3.2. Generate Selection Information System (SIS)

The Selection Information System is a 5-tuple expressed as

$$\mathbf{SIS} = \langle \mathbf{U}, \mathbf{X}, \mathbf{R}, \mathbf{V}_{\mathbf{X_j}}, \mathbf{h}_{\mathbf{X_j}} \rangle \tag{11}$$

where $U = [u_1, u_2, \ldots u_{n+n'}]$ is training dataset consisting of 'n' normal and 'n″' abnormal objects such that each object $U_i = \{X_i, R_i\}$. $X = [x_1, x_2 \ldots, x_p]$ is non-empty set of system variables and $R = \{r_1, r_2\}$ describes binary system categories.

The training dataset can be represented as

$$\mathbf{U} = \begin{bmatrix} x_{11} & x_{12} & . & x_{1p} & r_1 \\ x_{11} & x_{12} & . & x_{1p} & r_2 \\ . & . & . & . & . \\ x_{i1} & x_{i2} & . & x_{ip} & r_i \\ . & . & . & . & . \\ x_{n+n'1} & x_{n+n'2} & . & x_{n+n'p} & r_{n+n'} \end{bmatrix} \tag{12}$$

$\mathbf{V}_{\mathbf{X_j}}$ is set of values of any variable, $X_j \in \mathbf{X}$. Continuous variables are discretized and every variable can attain only a discrete set of values, $\mathbf{V}_{X_j} = \{V^1_{X_j}, V^2_{X_j} \ldots V^{m_j}_{X_j}\}$, where m_j is the number of discrete class labels or values of feature X_j. h_{X_j} is a function which maps objects from $\mathbf{U}$ to $\mathbf{V}_{\mathbf{X_j}}$ i.e. $h_{X_j} : \mathbf{U} \to \mathbf{V}_{\mathbf{X_j}}$. The function h_R determines the system state related to an object u_i i.e. $h_R(ui) = R_i$.

3.3.3. Determine equivalent classes for SSVs

Equivalent classes are generated based on the concept of indiscernibility. Indiscernibility implies that two objects u_i and $\mathrm{u_k}$ are similar or equivalent to each other with respect to a Subset of System Variables, $\mathbf{SSV} \subseteq \mathbf{X}$, if they cannot be distinguished based on their values for every variable $X_j \in \mathrm{SSV}$. An equivalent class obtained from **SSV** can be defined as

$$\mathbf{E}_{\mathrm{SSV},\boldsymbol{t}} = \{(u_i, u_k) \in \boldsymbol{U} \times \boldsymbol{U} | h_{X_j}(u_i) = h_{X_j}(u_k) \forall X_j \in \mathrm{SSV}\} \tag{13}$$

In other words, an equivalent class is a set of objects having same value for all the variables in a subset, $\mathbf{SSV} \subseteq \mathbf{X}$. The equivalence relation between two objects $\mathrm{u_i}$ and $\mathrm{u_k}$ can be represented as $h_{X_j}(u_i) = h_{X_j}(u_k)$ for all variables $X_j \in \mathbf{X}$. Further, the set **SSV** can create multiple equivalent classes depending on the various variable–value combinations existing in **U**. A set of multiple equivalent classes generated by **SSV** is given by

$$\boldsymbol{E_{SSV}} = \{\boldsymbol{E_{SSV}}_{,1}, \boldsymbol{E_{SSV}}_{,2} \ldots \boldsymbol{E_{SSV}}_{,T}\} \tag{14}$$

where T is the number of equivalent classes obtained for $\mathbf{SSV} \subseteq \mathbf{X}$.

An equivalent class ($\mathbf{E_{SSV,t}}$) can be described as combination of attribute–value pairs joined by the logical AND ($\wedge$) operator. Such a combination of attribute–value pairs is called a formula and is given by

$$\phi_{E_{SSV,t}} : (X_1 = V_{X_1}) \wedge (X_2 = V_{X_2}) \wedge \ldots \wedge (X_{|SSV|} = V_{X_{|SSV|}}) \tag{15}$$

Besides, all members of $\mathbf{E_{SSV,t}}$ can be related to exactly one of system states R_1 or R_2. In this case, the equivalent class is '*discriminative*' as its members belong to one and only one of the system states and applying the corresponding formula $\phi_{E_{SSV,t}}$ an unambiguous prediction of system state can be made for an object whose system state or category is not known. On the other hand, if few members of an equivalent class belong to state R_1 while the rest follow R_2, then the equivalent class is '*non-discriminative*' and a definite prediction of state system using the corresponding formula is not possible for an object whose system state is unknown. Further, the formulae are used to represent causal IF–THEN relationship between variables and predicted classes or system states. For example, the causal rule, **IF** $\boldsymbol{\phi}_{\mathbf{E_{SSV,t}}}(\mathbf{u_i}) \to$ **THEN** $\mathbf{R_1}$ indicates that if an object u_i satisfies the formula $\phi_{E_{SSV,t}}$ then its predicted state is R_1. Formulae corresponding to '*non-discriminative*' equivalent classes cannot definitively predict a system state for an object. In general an IF–THEN rule can be represented as

$$\psi_{E_{SSV,t}}: \text{ IF } \phi_{E_{SSV,t}}(u_i) \text{ THEN } R_1; \text{ IF } \phi_{E_{SSV,t}}(u_i) \text{ THEN } R_2; \text{ IF } \phi_{E_{SSV,t}}(u_i) \text{ THEN } R_1 or R_2 \quad (16)$$

IF–THEN rules like these provide information about underlying causal link between variables and system states.

3.3.4. Calculate degree of dependency $\delta_{SSV}(\mathbf{R})$

The degree of dependency measures the capability of the set of system variables **SSV** $\subseteq$ **X** to distinguish between the different categories **R** = {R_1, R_2}. Based on **SSV**, equivalent classes $\mathbf{E_{SSV}} = \{\mathbf{E_{SSV,1}}, \mathbf{E_{SSV,2}} \ldots \mathbf{E_{SSV,T}}\}$ are generated. Each $\mathbf{E_{SSV,t}}$ is associated with one or both of the categories which can be represented as follows,

$$\boldsymbol{R}(\boldsymbol{ESSV}, \boldsymbol{t}) = \bigcup_{u_i \in E_{SSV,t}} \{h_R(u_i)\} \quad (17)$$

The number of categories associated with $\mathbf{E_{SSV,t}}$ is given by

$$C(\boldsymbol{E_{SSV,t}}) = |R(\boldsymbol{E_{SSV,t}})| \quad (18)$$

An equivalent class $\mathbf{E_{SSV,t}}$ is *discriminative* if it is associated with one and only one category else it is *non-discriminative*. For **SSV** $\subseteq$ **X** the set of discriminative equivalent classes is given by

$$\boldsymbol{D}(\boldsymbol{E_{SSV}}) = \bigcup_{\boldsymbol{E_{SSV,t}} \in \boldsymbol{E_{SSV}}} \{\boldsymbol{E_{SSV,t}} | \boldsymbol{C}(\boldsymbol{E_{SSV,t}}) = 1\} \quad (19)$$

The dependency of system categories **R** on Subset of System Variables **SSV** is a measure of the capability of equivalent classes $\boldsymbol{E_{SSV,t}} \in \boldsymbol{E_{SSV}}$ to uniquely determine one and only one system category. Based on this notion, the dependency $\delta_{SSV}(R)$ is obtained as

$$\delta_{SSV}(\boldsymbol{R}) = \frac{|\boldsymbol{D}(\boldsymbol{E}_{SSV})|}{|\boldsymbol{U}|} \quad (20)$$

The dependency of system categories on all system variables **X** is given by

$$\delta_X(\boldsymbol{R}) = \frac{|\boldsymbol{D}(\boldsymbol{E_X})|}{|\boldsymbol{U}|} \quad (21)$$

3.3.5. Apply Genetic Algorithm to identify optimal SSV ($\boldsymbol{SSV_{opt}}$):

Genetic Algorithm (GA) is applied to generate $\mathbf{SSV_{opt}}$ which maximizes degree of dependency between system categories and variables. Candidate solutions **SSV** $\subseteq$ **X** are evaluated based on the fitness function

$$g(\boldsymbol{SSV}) = (p - |\boldsymbol{SSV}|)/p + min[\varepsilon, \delta_{SSV}(R)] \quad (22)$$

The algorithm searches for solutions which maximizes g(**SSV**). The first part of the fitness function $(p\text{-}\boldsymbol{SSV}|)/p$ ensures that candidate solutions with smaller cardinality get higher fitness score. The fraction of variables screened or eliminated is $(p\text{-}|\boldsymbol{SSV}|)/p$. This factor is equivalent to *regularization* of L_0 norm which acts as a penalty on number of variables included in building the MTS classifier. *Regularization* via reducing the fraction of variables screened allows feature selection to be solved as a joint maximization problem. The second part is a measure of goodness of the model built from the given subset of variables. The model-fit factor is set to a maximum hitting fraction ε so that all subsets of variables with *degree of dependency* more the cutoff (ε) gets the same score. Maximum hitting fraction is based on the degree of dependency of the full set of system variables and is determined by $\varepsilon = 0.9\delta_{\mathbf{X}}(\mathbf{R})$. Parameters used to run the GA are as follows:

- Chromosome population size = 100
- Number of generations = 50
- Probability of crossover = 0.30
- Probability of mutation = 0.05

Since GA gives more than one solution, the one which gives maximum accuracy of classification is selected. After the final subset of variables is identified, all equivalent classes related to the selected features are determined using Eqs. (13) and (14). Each equivalent class generates a corresponding formula as given by equation 15. From each formula a *'discriminative'* or *'non-discriminative'* IF–THEN rule linking one or both system states (R_1 or R_2 or both) is extracted as shown in Eq. (16).

3.4. Step IV: Rebuild predictor function and output class labels

Mahalanobis Scale is rebuilt with sample mean, standard deviations and correlation matrix based on $\mathbf{SSV_{opt}}$. Next, optimal threshold MD is obtained and based on this the percentage misclassification is calculated.

4. Case Studies

The proposed methodology is applied to 5 case studies each having binary system categories or decision classes. Table 3 provides a summary of the case studies in terms of the following: (i) number of system variables; (ii) system categories; (iii) size of training dataset; and (iv) size of test dataset. For all case studies, feature selection and development of the MTS classification model is done on the training dataset. The final MTS classifier, which is based on selected variables, is applied on the test dataset and percentage of correct classification is calculated as indicator of prediction accuracy. Performance is of the proposed MTS classifier is then compared with results from other classification techniques (Tables 5, 8, 9 and 11).

Table 3
Summary of case studies.

Case study	No. of system variables	System categories (R)	Size of training data	Size of test data
Breast cancer	9	R = {benign, malignant}	Benign – 20 Malignant – 10 Total –30	Benign – 444 Malignant – 239 Total –683
Credit card approval –I	15	R = {accepted, rejected}	Accepted – 45 Rejected –55 Total –100	Accepted – 383 Rejected –307 Total –690
Credit approval –II	14	R = {accepted, rejected}	Accepted – 46 Rejected – 54 Total – 100	Accepted – 228 Rejected – 204 Total – 432
Credit card approval –III	20	R = {accepted, rejected}	Accepted – 69 Rejected – 31 Total – 100	Accepted – 1399 Rejected – 601 Total – 2000
MONK's problem	6	R = {0,1}	0s – 30 1s – 20 Total – 50	0s – 30 1s – 20 Total – 50

Table 4
System variables in breast cancer case study.

Serial no.	Attributes	Range of values
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	System states or categories	Benign/malignant (B/M)

Data related to all the case studies is available in the digital repository for machine learning hosted by University of California, Irvine, United States (Bache & Lichman, 2013). Detail information related to the case studies such as names and locations have been kept anonymous. Actual variable names are known only for the breast cancer case-study. Hence, for demonstration, the causal IF–THEN rules, which are extracted from the final selected variables, are presented only for the breast cancer case study (Tables 6 and 7).

The steps of the methodology are illustrated in details using breast cancer case study and performance assessment is provided for all case studies. System variables of breast cancer case study and their range of values are listed in Table 4.

4.1. Illustrative Example – Breast Cancer case study

The steps of the methodology are illustrated in details as follows:

Step (I). Mahalanobis Distance (MD): Based on full set of variables and training data, the MD values are calculated using sample mean $\overline{X}$, Standard deviation SD, and correlation matrix C. The average of the MD values of 'benign' cases is 0.97, which is close to 1 and that of the 'malignant' cases is 73.86.

Step (II). Optimal threshold, $\text{MD}_{\text{T:}}$ Based on minimum misclassification, $\text{MD}_\text{T} = 28$ and class labels are (i) $\text{MD}(x) \leqslant 28 \rightarrow \text{CL}_1$; (ii) $\text{MD}(x) > 28 \rightarrow \text{CL}_2$. An objects x is misclassified using the current Mahalanobis Space if $x \in \text{CL}_1$ but $x \in R = \{benign\}$ according to training data or if $x \in \text{CL}_2$ *but* $x \in R = \{malignant\}$.

Step (III). Optimal Subset of System Variables (**$\mathbf{SSV_{opt}}$**): The steps to identify **$\mathbf{SSV_{opt}}$** for this case study is as follows:

A. Discretize continuous variables: The 9 system variables are continuous. Both Equal Frequency Binning (EFB) and Khiops Algorithm (KA) are used to discretize continuous variables in training data.

B. SIS: Two Selection Information Systems SIS-EFB and SIS-KA are created based discretized data from both algorithms. Overall the two approaches are termed as **MD-RST-EFB** and **MD-RST-KA**.

C. Equivalent classes: Based on the two Selection Information Systems, equivalent classes are generated.

Table 6
Breast cancer case study: IF–THEN rules based on three critical features selected in MTS-RST-EFB method.

Rule no.	Rule description
1	IF Clump Thickness = [5,6) AND Uniformity of Cell Size = [1,2) AND Bare Nuclei = [1, 2) THEN B[a]
2	IF Clump Thickness = [5, 6) AND Uniformity of Cell Size = [1, 2) AND Bare Nuclei = [2, 7) THEN B
3	IF Clump Thickness = [1, 5) AND Uniformity of Cell Size = [1, 2) AND Bare Nuclei = [1, 2) THEN B
4	IF Clump Thickness = [1, 5) AND Uniformity of Cell Size = [1, 2) AND Bare Nuclei = [2, 7) THEN B
5	IF Clump Thickness = [5, 6) AND Uniformity of Cell Size = [2, 6) AND Bare Nuclei = [1, 2) THEN B
6	IF Clump Thickness = [1, 5) AND Uniformity of Cell Size = [2, 6) AND Bare Nuclei = [1, 2) THEN B
7	IF Clump Thickness = [1, 5) AND Uniformity of Cell Size = [2, 6) AND Bare Nuclei = [2, 7) THEN B
8	IF Clump Thickness = [6, 10] AND Uniformity of Cell Size = [6, 10] AND Bare Nuclei = [2, 7) THEN M[b]
9	IF Clump Thickness = [6, 10] AND Uniformity of Cell Size = [6, 10] AND Bare Nuclei = [7, 10] THEN M
10	IF Clump Thickness = [5, 6) AND Uniformity of Cell Size = [6, 10] AND Bare Nuclei = [2, 7) THEN M
11	IF Clump Thickness = [5, 6) AND Uniformity of Cell Size = [2, 6) AND Bare Nuclei = [7, 10] THEN M
12	IF Clump Thickness = [6, 10] AND Uniformity of Cell Size = [2, 6) AND Bare Nuclei = [7, 10] THEN M
13	IF Clump Thickness = [1, 5) AND Uniformity of Cell Size = [2, 6) AND Bare Nuclei = [7, 10] THEN M
14	IF Clump Thickness = [6, 10] AND Uniformity of Cell Size = [1, 2) AND Bare Nuclei = [2, 7) THEN M
15	IF Clump Thickness = [6, 10] AND Uniformity of Cell Size = [2, 6) AND Bare Nuclei = [2, 7) THEN M

[a] B indicates benign.
[b] M indicates malignant.

Table 5
Comparison of results from breast cancer case study on test dataset. Bold values highlights results from proposed methodology.

Serial no.	Algorithm	Accuracy of classification (average)	Size of selected set of variables
1	**MTS-RST-KA**	**97.22**	**4**
2	**MTS-RST-EFB**	**97.1**	**3**
3	MTS-PSO	97.1	3
4	3-NN Stand Manhattan	97.0	5
5	Bayes (Pair wise dependent)	96.9	Variables selection is not done
6	Fisher linear discriminant analysis	96.8	9
7	Feature space mapping (FSM)	96.5	5
8	Naïve Bayes	96.4	
9	C4.5 (Decision Tree)	96.1	6
10	Only MD	95.9	Variables selection is not done
11	Ant colony optimization	95.47	9
12	MTS with OA	95.6	6
13	RIAC	95.0	Variables selection is not done
14	Neural network	94.15	3

Table 7
Breast cancer case study: IF–THEN rules based on three critical features selected in MTS-RST-KA method.

Rule no.	Rule description
1	IF Uniformity of Cell Size = [1, 2) AND Uniformity of Cell Shape = [1, 2) AND Marginal Adhesion = [2, 5) AND Bare Nuclei = [1, 2) THEN B[a]
2	IF Uniformity of Cell Size = [1, 2) AND Uniformity of Cell Shape = [2, 5) AND Marginal Adhesion = [5, 10] AND Bare Nuclei = [2, 7) THEN B
3	IF Uniformity of Cell Size = [1, 2) AND Uniformity of Cell Shape = [1, 2) AND Marginal Adhesion = [1, 2) AND Bare Nuclei = [1, 2) THEN B
4	IF Uniformity of Cell Size = [1, 2) AND Uniformity of Cell Shape = [1, 2) AND Marginal Adhesion = [1, 2) AND Bare Nuclei = [2, 7) THEN B
5	IF Uniformity of Cell Size = [1, 2) AND Uniformity of Cell Shape = [1, 2) AND Marginal Adhesion = [5, 10] AND Bare Nuclei = [1, 2) THEN B
6	IF Uniformity of Cell Size = [1, 2) AND Uniformity of Cell Shape = [2, 5) AND Marginal Adhesion = [1, 2) AND Bare Nuclei = [1, 2) THEN B
7	IF Uniformity of Cell Size = [1, 2) AND Uniformity of Cell Shape = [2, 5) AND Marginal Adhesion = [2, 5) AND Bare Nuclei = [1, 2) THEN B
8	IF Uniformity of Cell Size = [2, 6) AND Uniformity of Cell Shape = [2, 5) AND Marginal Adhesion = [2, 5) AND Bare Nuclei = [1, 2) THEN B
9	IF Uniformity of Cell Size = [2, 6) AND Uniformity of Cell Shape = [1, 2) AND Marginal Adhesion = [1, 2) AND Bare Nuclei = [1, 2) THEN B
10	IF Uniformity of Cell Size = [2, 6) AND Uniformity of Cell Shape = [1, 2) AND Marginal Adhesion = [1, 2) AND Bare Nuclei = [2, 7) THEN B
11	IF Uniformity of Cell Size = [6, 10] AND Uniformity of Cell Shape = [5, 10] AND Marginal Adhesion = [5, 10] AND Bare Nuclei = [2, 7) THEN M[b]
12	IF Uniformity of Cell Size = [6, 10] AND Uniformity of Cell Shape = [5, 10] AND Marginal Adhesion = [5, 10] AND Bare Nuclei = [7, 10] THEN M
13	IF Uniformity of Cell Size = [6, 10] AND Uniformity of Cell Shape = [2, 5) AND Marginal Adhesion = [5, 10] AND Bare Nuclei = [2, 7) THEN M
14	IF Uniformity of Cell Size = [2, 6) AND Uniformity of Cell Shape = [5, 10] AND Marginal Adhesion = [2, 5) AND Bare Nuclei = [7, 10] THEN M
15	IF Uniformity of Cell Size = [6, 10] AND Uniformity of Cell Shape = [5, 10] AND Marginal Adhesion = [2, 5) AND Bare Nuclei = [7, 10] THEN M
16	IF Uniformity of Cell Size = [2, 6) AND Uniformity of Cell Shape = [2, 5) AND Marginal Adhesion = [1, 2) AND Bare Nuclei = [7, 10] THEN M
17	IF Uniformity of Cell Size = [2, 6) AND Uniformity of Cell Shape = [2, 5) AND Marginal Adhesion = [5, 10] AND Bare Nuclei = [7, 10] THEN M
18	IF Uniformity of Cell Size = [1, 2) AND Uniformity of Cell Shape = [2, 5) AND Marginal Adhesion = [1, 2) AND Bare Nuclei = [2, 7) THEN M
19	IF Uniformity of Cell Size = [2, 6) AND Uniformity of Cell Shape = [5, 10] AND Marginal Adhesion = [1, 2) AND Bare Nuclei = [2, 7) THEN M

[a] B indicates benign.
[b] M indicates malignant.

D. Degree of dependency: The degree of dependency of full set of variables for the two cases is: (i) MD-RST-EFB; and (ii) MD-RST-KA. This is calculated implicitly by Rosetta and there is no such explicit value.

E. Genetic Algorithm based optimal optimal Subset of System Variables generation: The best $\mathbf{SSV_{opt}}$ generated is: (i) by MD-RST-EFB:{Clump Thickness, Uniformity of Cell Size, Bare Nuclei} & (ii) MD-RST-KA:{Clump Thickness, Uniformity of Cell Size, Marginal Adhesion and Bare Nuclei}. IF-THEN rules are extracted based on the selected variables. Table 6 and 7 enumerates the rules obtained from the results of MD-RST-EFB and MD-RST-KA respectively. In both cases, all IF–THEN rules are *'discriminative'* or indicates one and only one of the illness conditions – Benign (B) or Malignant (M).

Step (IV). Revise MDs and MD_T based on $\mathbf{SSV_{opt}}$: Using variables in $\mathbf{SSV_{opt}}$, the average of the MD values of 'benign' cases is 0.95, which is close to 1 and that of the 'malignant' cases is 61.34. MD_T based on $\mathbf{SSV_{opt}}$ is 1.1. Also, percentage misclassification using $\mathbf{SSV_{opt}}$ is (i) for training data: 3.33%; and (ii) for test data: 2.78%.

The plot of MD values of training dataset based on $\mathbf{SSV_{opt}}$ is shown in the Fig. 3. Similar results for test dataset are presented in Fig. 4.

The comparison of performance is done based on two criteria – (i) percentage accuracy of classification; (ii) size of selected set of variables.

Results are given in descending order of accuracy of classification. Table 5 compares the results from Breast Cancer case study.

4.2. Results from additional case studies

For the remaining case studies, Tables 8–12 show comparison of results.

4.2.1. Credit Card Approval case study-I

This dataset pertains to decision on credit card application. It consists of 9 discrete and 6 continuous variables with two target classes – *accepted* and *rejected*. The results are presented in Table 8 which shows that MTS-RST-KA gives second best accuracy of classification among other methods.

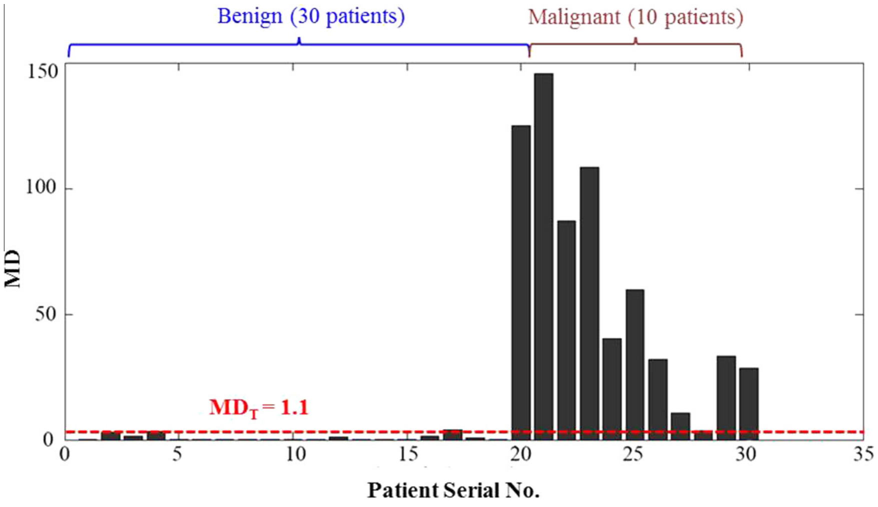

Fig. 3. Plot of MD values for training dataset based on $\mathbf{SSV_{opt}}$.

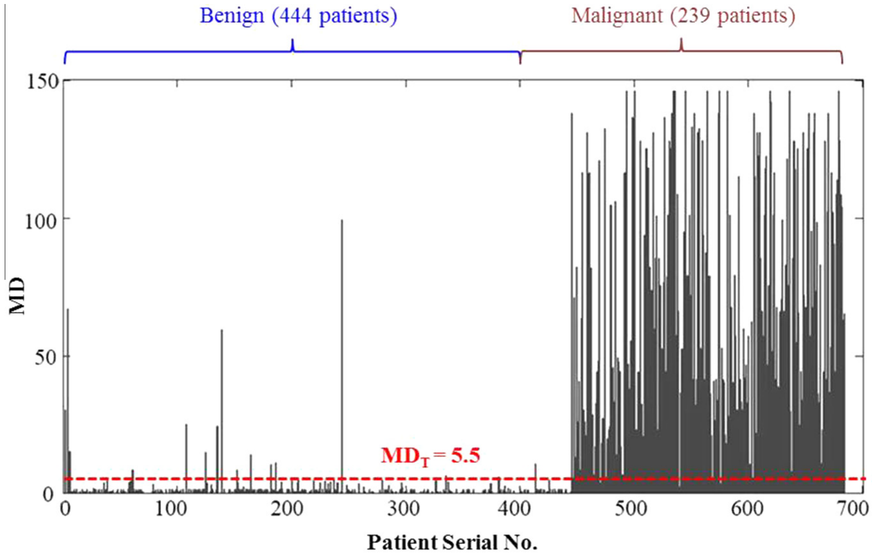

Fig. 4. Plot of MD values for test dataset based on $\mathbf{SSV_{opt}}$.

Table 8
Comparison of results from credit card approval case study – I on test dataset. Bold values highlights results from proposed methodology.

Serial no.	Algorithm	Accuracy of classification
1	MTS-PSO	98.9
2	**MTS-RST-KA**	**87.3**
3	**MTS-RST-EFB**	**85.2**
4	Pessimistic pruning	83.6
5	Cost complexity pruning	82.9
6	Reduced error pruning	82.6
7	Production rules	82.2
8	Decision tree	79
9.	Only MD	Singular covariance matrix obtained

Table 9
Comparison of results from credit card approval case study-II on test dataset. Bold values highlights results from proposed methodology.

Serial no.	Algorithm	Accuracy of classification
1.	Clustering GP	86.3
2.	**MTS-RST-EFB**	**85.5**
3.	**MTS-RST-KA**	**85.4**
4.	Frontier-based tree pruning algorithm	85.4
5.	Cost complexity pruning	85.16
6.	MTS-PSO	85.10
7.	Decision Tree – C4.5	84.1
8.	Extended separate-and-conquer propositional rule induction algorithm	80.6
9.	Simple genetic programming	78.0
10.	Only MD	76.81

4.2.2. Credit Card Approval case study-II

The dataset contains the credit card approval for 690 customers. The objective is to identify whether a credit card application is approved or not. The results are presented in Table 9 where MTS-RST-KA stands third and with MTS-RST-KA second and clustering GP outperforming all the other methods.

Table 10
Comparison of results from credit card approval case study-III on test dataset. Bold values highlights results from proposed methodology.

Serial no.	Algorithm	Accuracy of classification
1.	Simple GP	72.9
2.	Decision Tree – C4.5	72.8
3.	**MTS-RST-KA**	**72.6**
4.	**MTS-RST-EFB**	**72.6**
5.	MTS-PSO	72.45
6.	Clustering genetic programming	72.2
7.	Substructure discovery using minimum description length concept learner	71.52
8.	MD only	70.85
9.	First order inductive learner	70.66
10.	Extended separate-and-conquer-propositional-rule induction algorithm	70.5

Table 11
Comparison of results from *MONK's problem case study* on test dataset. Bold values highlights results from proposed methodology.

Serial no.	Algorithm	Accuracy of classification
1	Decision Tree – C4.5	97.2
2	**MTS-RST-KA**	**97.0**
3	MTS-PSO	97.0
4	**MTS-RST-EFB**	**97.0**
5	Feature space mapping, fuzzy rules	95.5
6	Incremental decision tree model	95.2
7	Real time dynamic programming using belief states	95.16
8	Iterative Dichotomizer 3	94.4
9	Multi-layer perceptron	93.1
10	Only MD	81.71

4.2.3. Credit Card Approval case study-III

This is another case study on credit card approval. Results are shown in Table 10. The performance of all the algorithms is close to each other with MTS-RST-KA being third best.

Table 12
Percentile ranks of algorithms based on accuracy of classification. Bold values highlights results from proposed methodology.

Method	Case study					
	Breast cancer	Credit card approval-I	Credit card approval-II	Credit card approval-III	MONK's problem	**Mean**
MTS-RST-KA	**96.2**	**81.25**	**83.33**	**72.2**	**94.44**	**85**
MTS-RST-EFB	**84.6**	**68.75**	**83.33**	**66.67**	**72.22**	**75**
MTS-PSO	84.6	93.75	38.89	50.0	72.22	69
MTS-OAs	26.9	X[a]	X	X	X	X
MD only	95.4	X	76.81	70.85	81.71	81
Decision Tree	34.6	6.25	27.78	83.33	94.44	49
Simple GP	X	X	5.56	94.44	X	50
Clustering GP	X	X	94.44	38.89	X	67
FSM	50.00	X	X	X	50.0	50

[a] X indicates results are not available either because limitations of the method or non-availability of results in literature for the case.

4.2.4. MONK's Problem case study

This case is related to classification of objects as 0 or 1. The results are presented in the Table 11 where the second best performance is achieved by MTS-RST-KA, being slightly lower than Decision Trees C4.5.

4.3. Performance Assessment

Assessment of performance of the proposed methodology is done based on variance of the percentile rank obtained by an algorithm in a case study. The percentile rank of an algorithm in a case study is calculated from accuracy of classification using

$$r = \frac{c_s + 0.5F_s}{N} \times 100 \tag{23}$$

where c_s is the count of all accuracy less than s, F_s is the frequency of accuracy of interest s and N is the number of algorithms examined for the case study. Table 12 shows the percentile ranks and their variance for the algorithms which were used at least twice in the case studies.

Fig. 5 shows Whisker charts of percentile ranks for 4 classifications methods whose results are available for all 5 case studies. MTS-RST-KA and MTS-RST-EFB perform consistently better compared to other methods as indicated by the smaller range in their corresponding Whisker charts.

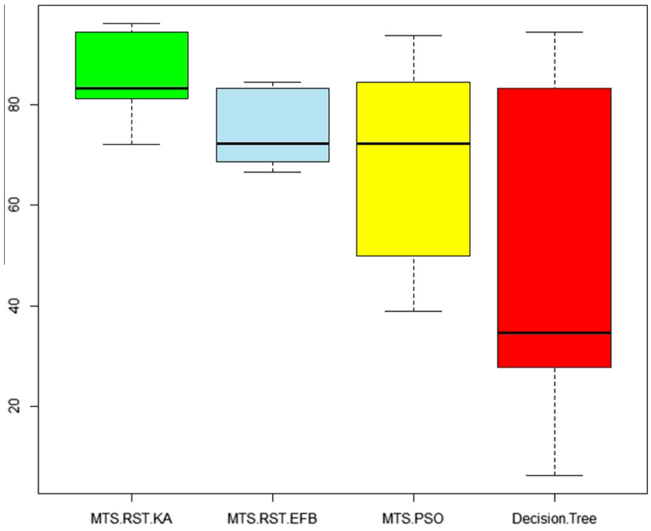

Fig. 5. Whisker charts of percentile ranks based on accuracy of classification.

5. Conclusion

Classification is an important topic in machine learning and is applied for drawing inference about unknown behavior and state of an object. Predicting the unknown state or behavior enables decision making, which makes classification techniques very practical tools. Many researches have sought Mahalanobis–Taguchi System (MTS) as a classification and decision-making tool for numerous applications. However, the there is scope to improve the classification performance of MTS and enhance its utility as a decision-making tool. This paper reviews related work on MTS, identifies the scope for improvement and develops a methodology and integrates the proposed solution with the overall existing framework of MTS to build an enhanced MTS classifier. In Section 5.1, the contribution of the current research in data-mining focusing on developing an enhanced MTS classifier is summarized. The discussion also includes explanation of key ideas to provide a conceptual insight of the proposed methodology and the improved performance of the MTS classifier as an outcome. Secondly Section 5.2 describes potential future work, which has been identified by the current research.

5.1. Contribution to Mahalanobis–Taguchi System

Feature selection is an important part of MTS and in general of most classification methods. The process of feature selection aims to find a subset of variables which improves goodness-of-model, prevents over-fitting on training data and provides a comprehensive description of underlying process and behavior. The contribution of the proposed methodology covers all of these aspects:

I. *Measure for goodness-of-model*: Past research on MTS have applied either Signal-to-Noise ratios or Total Weighted Misclassification as measures of goodness-of-model. This paper presents *degree-of-dependency* or conditional probability of predicted states on variables to evaluate model fit. The basis for this measure is to explore a probabilistic mechanism of modeling the behavior or state with respect to known attributes.

II. *Avoiding over-fitting on training data*: This research introduces *regularization* or penalty for over-fitting in the feature selection process of MTS. The proposed *regularization* criteria fraction of variables screened ($\frac{p-|SSV|}{p}$) is equivalent to a penalty of L_0 norm. A joint maximization problem, which improves both goodness-of-model and lack of over-fitting, is proposed to select important variables. *Regularization* has significant effect on model performance especially on test dataset. This is shown by higher prediction accuracy by the proposed variants of MTS namely, MTS-RST-EFB and MTS-RST-KA, as compared to previously studied forms such as MTS-Orthogonal Arrays and MTS-Particle Swarm Optimization.

III. *Comprehensive description of underlying process*: IF–THEN rules are extracted from the variables selected by the proposed feature selection process to provide a description of the cause-effect relationship between the critical variables and the predicted classes or states. IF–THEN rules provide a comprehensive and intuitive way to represent causality and provide valuable insights to domain experts who rely on data-mining to extract hidden patterns and represent in a usable and informative format for further analysis using domain-specific tools. Including IF–THEN for the MTS classifier enhances its utility to a classification cum causality analysis tool.

5.2. Future Work

Potential future research is categorized under two topics:

(i) *Methods on Mahalanobis Distance (MD)-based classifiers*: Research under this topic can focus on applying or enhancing Mahalanobis Distance (MD) as the key classifying parameter in MTS. Another opportunity to explore further improvements in feature selection by either extending current feature selection paradigms or developing new ones which address peculiarities of the dataset such as multicolinearity, low number of observations to number of variables ratio (also known as curse of dimensionality). Measure of goodness-of-model and penalty for over-fitting will have to be customized to address the specific issues. Further, potential enhancement of current research can be done as future work to address multi-class classification problems having more than two categories. A potential approach for multi-category classification is marking a particular class as 1 and the rest as 0. This reduces the problem to binary classification for the class marked as 1. The process is then repeated for each category taken one at a time. Lastly, the feature selection process can also include cross-validation to create small test datasets from training dataset to check and optimize prediction accuracy during the process of model building.

(ii) *Methods on hybrid classifiers*: Ensemble is a powerful technique to achieve improvement in accuracy of predictive models, whereby individual methods, which do not perform consistently best in different problems and datasets, are brought together to provide predictions which are more accurate than those made by individual methods. Ensemble combines results from multiple methods via voting (in classification) and averaging (in regression) to produce more accurate predictions (Dietterich, 2000). As shown Table 12, methods like Decision Trees or Genetic Programming (GP) show huge variations in performance over the five case studies whereas MTS-RST-EFB, MTS-RST-KA show performance which is consistently better than average overall. A significant insight from the results is that every predictive model captures certain information from the data which may not be captured by other models. This inherent ability of different predictive models to learn different aspects the data is exploited in ensemble. Therefore, improvement in prediction of the proposed variants of MTS can be achieved by combining with other classification models such as Decision Trees, GP etc.

In summary, the current research has developed a new variant of MTS by proposing a feature selection method which explores a new measure of goodness-of-model in terms of conditional probability of system states on subset of variables. Additionally, penalty for over-fitting or *regularization* factor has been introduced in the feature selection process for the MTS classifier. The proposed variants of MTS namely MTS-RST-EFB and MTS-RST-KA show better classification performance than existing methods such as MTS-OA and MTS-PSO. Lastly, there is opportunity to combine the proposed and existing variants of MTS with other classification techniques via ensemble to further improve classification accuracy.

References

Abraham, B., & Variyath, A. M. (2003). Discussion. *Technometrics, 45*(1), 22–24.

Aggarwal, C. C., Hinneburg, A., & Keim, D. A. (2001). *On the surprising behaviour of distance metrics in high dimensional space*. Berlin Heidelberg: Springer, pp. 420–434.

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. In *ACM SIGMOD record* (Vol. 22, no. 2, pp. 207–216). ACM.

Ahn, H., Ahn, J. J., Byun, H. W., & Oh, K. J. (2011). A novel customer scoring model to encourage the use of mobile value added services. *Expert Systems with Applications, 38*(9), 11693–11700.

Ahn, H., Ahn, J. J., Oh, K. J., & Kim, D. H. (2011). Facilitating cross-selling in a mobile telecom market to develop customer classification model based on hybrid data mining techniques. *Expert Systems with Applications, 38*(5), 5005–5012.

Al-Aidaroos, K., Bakar, A. A., & Othman, Z. (2010). Data classification using rough sets and naïve Bayes. In *Rough set and knowledge technology* (pp. 134–142). Berlin Heidelberg: Springer.

Alpaydin, E. (2004). *Introduction to machine learning*. MIT press.

An, A., Huang, Y., Huang, X., & Cercone, N. (2005). *IFeature selection with rough sets for web page classification,Transactions on Rough Sets II*. Springer Berlin Heidelberg (pp. 1–13). Springer Berlin Heidelberg.

Antonie, M. L., Zaiane, O. R., & Coman, A. (2001). Application of data mining techniques for medical image classification. *MDM/KDD*, 94–101.

Bache, K. & Lichman, M. (2013). UCI machine learning repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Bell, D. A., Guan, J. W., & Liu, D. Y. (2005). Mining association rules with rough sets. In *Intelligent data mining* (pp. 163–184). Berlin Heidelberg: Springer.

Berry, M. J., & Linoff, G. (1997). *Data mining techniques: for marketing, sales, and customer support*. John Wiley & Sons Inc.

Boulle, M. (2004). Khiops: A statistical discretization method of continuous attributes. *Machine Learning, 55*(1), 53–69.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52*(3), 345–370.

Bredensteiner, E. J., & Bennett, K. P. (1998). Feature minimization within decision trees. *Computational Optimization and Applications, 10*(2), 111–126.

Chan, P. K., Fan, W., Prodromidis, A. L., & Stolfo, S. J. (1999). Distributed data mining in credit card fraud detection. *IEEE Intelligent Systems and their Applications, 14*(6), 67–74.

Changchien, S., & Lu, T. C. (2001). Mining association rules procedure to support online recommendation by customers and products fragmentation. *Expert Systems with Applications, 20*(4), 325–335.

Das, P., & 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.

Dietterich, T. (1995). Overfitting and undercomputing in machine learning. *ACM Computing Surveys (CSUR), 27*(3), 326–327.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Multiple classifier systems* (pp. 1–15). Berlin Heidelberg: Springer.

Duch, W., Adamczak, R., Grabczewski, K., & Zal, G. (1998). A hybrid method for extraction of logical rules from data. In *Second polish conference on theory and applications of artificial intelligence*, Łódź (pp. 61–82).

Durrant, R. J., & Kabán, A. (2010). Compressed fisher linear discriminant analysis: classification of randomly projected data. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1119–1128). New York: ACM.

Gamberger, D., Lavrac, N., & Jovanoski, V. (1999). High confidence association rules for medical diagnosis. *Proceedings of IDAMAP, 99*, 42–51.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research, 3*, 1157–1182.

Hamilton, H. J., Shan, N., & Cercone, N. (1996). RIAC: A rule induction algorithm based on approximate classification. Computer Science Department, University of Regina.

Hassanien, A. E., & Ali, J. M. (2004). Enhanced rough sets rule reduction algorithm for classification digital mammography. *Journal of Intelligent Systems, 13*(2), 151–171.

Jin, X., & 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.

Komorowski, J., & Øhrn, A. (1999). Modelling prognostic power of cardiac tests using rough sets. *Artificial Intelligence in Medicine, 15*(2), 167–191.

Kotsiantis, S., & Kanellopoulos, D. (2006). Discretization techniques: A recent survey. *GESTS International Transactions on Computer Science and Engineering, 32*(1), 47–58.

Lavrač, N. (1999). Selected techniques for data mining in medicine. *Artificial Intelligence in Medicine, 16*(1), 3–23.

Lee, Y. C., & 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.

Li, Y., Shiu, S. C. K., Pal, S. K., & Liu, J. N. K. (2006). A rough set-based case-based reasoner for text categorization. *International journal of approximate reasoning, 41*(2), 229–255.

Liu, H., Hussain, F., Tan, C. L., & Dash, M. (2002). Discretization: An enabling technique. *Data Mining and Knowledge Discovery, 6*(4), 393–423.

Malerba, D., Esposito, F., Lisi, F. A., & Appice, A. (2003). Mining spatial association rules in census data. Research in Official Statistics. v5 i1, 19–44.

Mannar, K., Ceglarek, D., Niu, F., & Abifaraj, B. (2006). Fault region localization: product and process improvement based on field performance and manufacturing measurements. *IEEE Transactions on Automation Science and Engineering, 3*(4), 423–439.

Myers, R. H. (1990). *Classical and modern regression with applications* (Vol. 2). Belmont, CA: Duxbury Press.

Nakayama, H., Hattori, Y., & Ishii, R. (1999). Rule extraction based on rough set theory and its application to medical data analysis, Proceedings of 1999 IEEE International Conference onSystems, Man, and Cybernetics, 5, 924–929.

Ng, A. Y. (2004). Feature selection, regularization, L 1 vs. L 2 and rotational invariance. In *Proceedings of the twenty-first international conference on machine learning* (pp. 78). ACM.

Department of Mathematics 2004, *Orthogonal arrays (Taguchi designs)*, University of York, Available from: http://www.york.ac.uk/depts/maths/tables/orthogonal.htm [May 2004].

Pawlak, Z., & Ceglarek, D. (1982). Rough sets. *International Journal of Computer & Information Sciences, 11*(5), 341–356. http://dx.doi.org/10.1016/j.procir.2013.07.067.

Pal, A., & Ceglarek, D. (2013). Modeling of decision making process for product service failure diagnosis. *Procedia CIRP, 11*, 32–37. http://dx.doi.org/10.1016/j.procir.2013.07.067.

Pal, A., & 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.

Parpinelli, R. S., Lopes, H. S., & Freitas, A. A. (2001). An ant colony based system for data mining: applications to medical data. In *Proceedings of the genetic and evolutionary computation conference (GECCO-2001)* (pp. 791–797).

Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research. arXiv preprint arXiv:1009.6119.

Rai, B. K., Chinnam, R. B., & 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.

Reséndiz, E., Moncayo-Martínez, L. A., & Solís, G. (2013). Binary ant colony optimization applied to variable screening in the Mahalanobis–Taguchi system. *Expert Systems with Applications, 40*(2), 634–637.

Reséndiz, E., & 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.

Schölkopf, B., & Smola, A. J. (2001). *Learning with kernels: Support vector machines, regularization, optimization, and beyond (adaptive computation and machine learning)*. MIT Press.

Setiono, R., & Liu, H. (1997). Neural-network feature selector. *IEEE Transactions on Neural Networks, 8*(3), 654–662.

Slowinski, R. (1992). Rough sets with strict and weak indiscernibility relations. In *IEEE international conference on fuzzy systems* (pp. 695–702). IEEE.

Song, H. S., & Kim, S. H. (2001). Mining the change of customer behavior in an internet shopping mall. *Expert Systems with Applications, 21*(3), 157–168.

Sun, C., Nguyen, K., Vu, L., & Bisland, S. C. (2006). Prognostic/diagnostic health management system (PHM) for fab efficiency. In *The 17th annual SEMI/IEEE advanced semiconductor manufacturing conference, ASMC 2006* (pp. 433–438). IEEE.

Taguchi, G., Chowdhury, S., & Wu, Y. (2005). *Taguchi's quality engineering handbook*. Wiley.

Taguchi, G., & Jugulum, R. (2002). *The Mahalanobis–Taguchi strategy, a pattern technology system*. Wiley. com.

Thangavel, K., & Pethalakshmi, A. (2009). Dimensionality reduction based on rough set theory: A review. *Applied Soft Computing, 9*(1), 1–12. http://dx.doi.org/10.1016/j.asoc.2008.05.006.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Woodall, W. H., Koudelik, R., Tsui, K. L., Kim, S. B., Stoumbos, Z. G., & Carvounis, C. P. (2003). A review and analysis of the Mahalanobis–Taguchi system. *Technometrics, 45*(1), 1–15.

Yang, T., & Cheng, Y. T. (2010). The use of Mahalanobis–Taguchi system to improve flip-chip bumping height inspection efficiency. *Microelectronics Reliability, 50*(3), 407–414.