*Send Orders for Reprints to reprints@benthamscience.ae*



BENTHAM OPEN

CrossMark

**The Open Cybernetics & Systemics Journal**

Content list available at: www.benthamopen.com/TOCSJ/

DOI: 10.2174/1874110X01610010250

RESEARCH ARTICLE

# A Risk Decision-making Approach to Customs Targeting

Zhi Xiao, Hongshan Xiao* and Yu Wang

*School of Economics and Business Administration, Chongqing University, Chongqing, 400030, China*

Received: February 26, 2016 Revised: September 28, 2016 Accepted: October 04, 2016

**Abstract:** This paper focuses on the risk decision-making problem in customs targeting, whose major responsibility is to inspect the smuggling goods in import/export declarations. In this problem, the estimated smuggling probabilities of import/export goods, which can be obtained by applying statistical analysis to observations (samples), are needed for accurate inspection decision. A critical presumption for statistical analysis is that the samples are homogeneous or subject to certain distributions. Therefore, clustering techniques are usually employed for preprocessing the samples. However, severe heterogeneity and abnormality exist among the large amount of samples and thus hinder the performance of conventional clustering methods for preprocessing. To deal with this problem, a dynamic *K*-means clustering method is developed in this paper. Through optimizing the validity function that indicates the goodness of clustering result, the entire samples are iteratively divided into a number of clusters. Based on the dynamic *K*-means clustering method and logistic regression, a risk decision-making approach is proposed and applied to China's customs targeting. The empirical results show that the proposed approach improves the accuracy and decreases the risk of inspection decision.

**Keywords:** Customs Targeting, Data Preprocessing, *K*-means Clustering, Logistic Regression, Risk Decision-Making.

## 1. INTRODUCTION

Risk decision-making problems exist widely in public administration, business administration, and engineering management. Traditionally, the expected utility model is a fundamental paradigm for risk decision-making problems [1 - 3]. Nevertheless, the rapid growth of data nowadays is changing this paradigm and confronting the decision-makers with new challenges. Therefore, statistical analysis has been employed in risk decision-making, and a number of researches have been carried out. Typical examples are credit scoring and classification [4 - 6], facility maintenance [7], medical and hospital practice [8, 9], intermittent demand of spare parts in manufacturing [10, 11], and emergency response [12, 13].

Customs targeting, whose major responsibility is to decide whether goods in an import/export declaration should be inspected, is a typical risk decision-making problem. In customs targeting, there are a huge amount of historical observations (samples) stored in the database. Each historical observation represents a declaration of goods, which consists of some attributes and a specific state of nature. One state of nature is that goods match the regulations on customs declaration (this usually means legality), and the other state of nature is that goods do not match the regulations on customs declaration (this usually means smuggling). For a newly encountered declaration of goods, the customs officials do not know exactly whether it is of smuggling before inspection, and they need to estimate its smuggling probability based on the historical observations (samples) in order to select the optimal action (inspection or no-inspection). Therefore, statistical analysis can be applied for probability estimation.

For efficient statistical analysis, it is required that the samples are homogeneous or subject to some certain distributions [14]. To ensure the samples' homogeneity, clustering is usually adopted as the preprocessing technique. Through clustering, a data set is partitioned into a number of groups, so that samples in the same group are as similar as

* Address correspondence to this author at the Chonqging University, 174# Shazhengjie Street, Shapingba District, Chongqing, 400030, China; Tel: +86-23-65105761; Fax: 86-23-65105761; E-mail: xiaohongshan@cqu.edu.cn

possible, and are as dissimilar as possible from the samples in other groups [15]. Smith *et al*. [16] employ the clustering techniques as the data preprocessing method and apply logistic regression model to estimate the probabilities that policy holders renew and terminate their policies. In electrical load forecasting, Sfetsos [17] proposes a hybrid algorithm in which the *K*-means clustering is employed to generate clusters of data whose characteristics are similar, so that they could be described by the same linear model. Pavlidis *et al*. [18] employ the clustering techniques to identify neighbourhoods in the reconstructed state space of the system in one-step prediction for daily foreign exchange spot rate. Hua *et al*. [14] propose a risk decision-making approach based on *K*-means clustering and logistic regression for China's customs inspection decision. These researches confirm the effectiveness of clustering for improving the prediction accuracy. However, most of the applied *K*-means clustering methods are "static", which means that the number of clusters is predefined and constant in the clustering process.

In China's customs administration, declared import/export goods are put into more than eight thousand merchandise categories due to the variety and diversity of import/export trades. Before statistical analysis, goods in these merchandise categories should be clustered into a number of groups in each of which the observations are approximately homogeneous. Nevertheless, when the number of samples is large with severe heterogeneity, conventional *K*-means clustering method is difficult to apply, or has little effect on improving the results of statistical analysis. For examples, a piece of metal equipment may have a net weight of more than one hundred thousand kilograms, while a lampshade has a unit weight of less than one kilogram; the unit price of diode is less than one dollar, and the unit price of an automatic numerical control system of printing machine is more than one million dollars. Since most samples have a relatively small or medium range of variable values, the conventional *K*-means clustering techniques will put them into one cluster. In addition, the abnormally large cluster could not be easily divided into smaller sub-clusters by simply increasing the number of clusters. Consequently, the performance of clustering for data preprocessing is hindered.

To deal with this problem, a dynamic *K*-means clustering method is developed to preprocess the samples before statistical analysis. In comparison with 'static' clustering paradigm, 'dynamic' means that the number of clusters varies according to the optimal validity function in the clustering process. As for the clustering methods with dynamic characteristics, Bhargavi and Gowda [19] develop a clustering validity index to dynamically terminate the clustering process. Benítez *et al*. [20] propose a dynamic clustering segmentation algorithm to profiling of energy load. Ozturk *et al*. [21] use the Artificial Bee Colony (ABC) algorithm to optimize the clustering validity function, without determining the number of clusters in advance. The proposed dynamic *K*-means clustering method differs from those in the literature in two aspects. First, it searches for the optimal validity function in an iterative manner, and may get a better validity function given a specific number of clusters. Second, it is able to handle samples with severe heterogeneity and various intervals of values, which would be demonstrated in the case study.

In the proposed dynamic *K*-means clustering method, a validity function is defined to evaluate the performance of clustering result, and the obtained clusters are divided into sub-clusters iteratively according to their intra-cluster characteristics. The main purpose of this method is to optimize the clustering validity function, as well as to deal with the problem that most samples would form into one cluster and significantly impact the clustering result. It has been found that for the same number of clusters obtained, the proposed dynamic *K*-means clustering method outperforms the conventional *K*-means clustering method with respect to the optimization of validity function. Therefore, it is effective for preprocessing large amount of samples with severe heterogeneity.

After data preprocessing, we use logistic regression, a well-known statistical analysis technique for estimating the occurrence probability of a state of nature, to predict the occurrence probability of smuggling. Logistic regression is a powerful and time-efficient technique for estimating the probability of occurrence of an event, especially when there are huge amount of data. Also, it is able to handle categorical or nominal attributes, which is quite suitable for our task. Until recently, logistic regression has been widely applied to risk decision-making problems in medical, social and economic researches. For examples, DesJardins [22] proposes a logistic regression model to evaluate the probability that a student enrolls after being accepted. Lowe and Parvar [23] present a logistic regression model for the bid decision process. The empirical results show that the model is effective in predicting bid/no-bid decisions. Pourahmad *et al*. [24] develop a fuzzy logistic regression model to detect the SLE disease. Lacagnina *et al*. [25] study a fuzzy model, which is expressed in rule-form, and logistic regression for diagnostic decision making in patients with chronic nasal symptoms. Based on the dynamic *K*-means clustering method and logistic regression, a risk decision-making approach is proposed and applied to China's customs targeting. The empirical results show that the proposed approach could improve the accuracy and decrease the risk of inspection decision.

The rest of this paper is organized as follows. In Section 2, the proposed risk decision-making approach based on dynamic *K*-means clustering and logistic regression is elaborated. Section 3 reports the case study and application results of the proposed approach to China's customs targeting. In the last section, the paper ends with some concluding remarks.

## 2. THE RISK DECISION-MAKING APPROACH

In this section, a risk decision making approach based on the dynamic *K*-means clustering and logistic regression is proposed. The dynamic *K*-means clustering method is used to divide the samples iteratively into groups, in each of which the samples are approximately homogeneous, and logistic regression is applied to each group of samples to obtain the estimated smuggling probabilities. After the above steps, the risk decision making rule is applied to make the optimal targeting decision. To make the procedures of the proposed risk decision-making approach more clear, we first briefly introduce the conventional *K*-means clustering and present the validity function.

### 2.1. *K*-means Clustering and Its Validity Function

*K*-means clustering is a popular algorithm particularly suited for partitioning large amount of samples [26]. The main process of the *K*-means clustering includes the following four steps [15]: (1) selection of the initial *K* centroids. In this step, *K* samples are often randomly selected as the centroid of *K* clusters; (2) assignment. In this step, each sample is assigned to the nearest cluster in sequence (the distance is usually measured in Euclidean distance metric). Meanwhile, the centroid of newly changed cluster should be updated; (3) computation of the quality function of clustering; (4) evaluation. If the quality function does not change, stop and output the final clustering results, else go to step (2).

In *K*-means clustering, an important problem is to determine the number of clusters, which is usually called the cluster validity [27]. In the literature, many cluster validity indexes such as partition coefficient (PC) and partition entropy (PE) [27], XB index [28], modified partition coefficient (MPC) [29], DB index and two of its generalized versions [30], SC index [31], Dunn's index [32], and SV index for crisp clustering and SVF index for fuzzy clustering [33], have been proposed to determine the optimal number of clusters and evaluate the fitness of partitions produced by clustering algorithm. In this paper, a new validity function for hard *K*-means clustering is proposed based on the idea of Wu and Yang [34], in which two main aspects of clustering validity are considered, i.e., compactness of each cluster and separation between clusters. However, in their work, the validity function is for soft *K*-means clustering. In order to fit it to hard *K*-means clustering for supervised clustering, the validity function is modified as follows.

Assume the data set $S$ consisting of $N$ samples $X_1, X_2,...,X_N$ (class labels are not considered in unsupervised clustering) is partitioned into $K$ clusters $C_1, C_2,...,C_K$ The centroids of these clusters are denoted by $r^1, r^2,...,r^K$, in which $r^K$ ($k = 1, 2,...,K$) is defined as

$$\mathbf{r}_0^k = \sum_{i=1}^{m_k} \mathbf{X}_i^k / m_k \tag{1}$$

where $m_k$ is the number of samples in cluster $C_k$, and $X_i^k$, ($i = 1, 2,..., m_k$) denotes the $i$th sample that belongs to the cluster $C_k$. The index of compactness of clusters obtained is defined as

$$Intra = \frac{1}{K}\sum_{k=1}^{K} Intra(k) = \frac{1}{K}\sum_{k=1}^{K}[\sum_{i=1}^{m_k} | \mathbf{X}_i^k - \mathbf{r}_0^k | /(m_k \max_{j\in[1,m_k]} | \mathbf{X}_j^k - \mathbf{r}_0^k |)] \tag{2}$$

In formula (2), the term $Intra(k)$ represents the relative distance between each sample and the centroid, with respect to the maximal distance between all samples and the centroid, in the $k$th cluster. It is defined as

$$Intra(k) = (\sum_{i=1}^{m_k} | \mathbf{X}_i^k - \mathbf{r}_0^k |) /(m_k \max_{j\in[1,m_k]} | \mathbf{X}_j^k - \mathbf{r}_0^k |) \tag{3}$$

Intuitively, a compact cluster requires the term *Intra* defined in formula (2) to be small. Since $0 < Intra(k) \leq 1$, it can be easily seen that $0 < Intra \leq 1$ . For clusters with only one sample, we define $Intra(k) = 1$ .

For a good clustering result, not only each cluster obtained is compact, but also these clusters are well separated from each other. For this purpose, the index of separation between clusters is defined as follows

$$Inter = \exp(-D/\beta), \tag{4}$$

in which $D = 2\sum_{1\le i<j\le K} |\mathbf{r}_0^i - \mathbf{r}_0^j| / [K(K-1)]$ is the average distance between each pair of cluster centroids, and $\beta = \sum_{k=1}^{K} |\mathbf{r}_0^k - \mathbf{r}_0| / K$ is the average distance of each cluster centroid $r^k$ to the centroid of all samples. Apparently, $0 < Inter \le 1$, and a smaller *Inter* indicates that clusters obtained are more separated from each other.

The objective of *K*-means clustering is to find out the optimal *K* clusters each of which is compact and separated from others. Therefore, the validity function considering both the compactness and separation of clusters is defined as follows

$$VF(K) = Intra \bullet Inter, \tag{5}$$

in which *Intra* and *Inter* are defined in formulas (2) and (4), respectively. Clearly, the smaller the $VF(K)$ is, the better the clustering results are.

Algorithm 1 The dynamic *K*-means clustering method

Input: $N$ samples in the data set.

Output: Clusters that contain all $N$ samples.

*Step* 1. Set $K_{\max} = \sqrt{N}$;

*Step* 2. Loop for $K$=2 to $K_{\max}$:

Use conventional *K*-means clustering to divide the $N$ samples into $K$ clusters $\{C_1, C_2, \cdots, C_K\}$;

*Step* 3. Compute the validity function values $VF(2), \cdots, VF(K_{\max})$ using Eq. (5), set $VF_{opt} = \min_{k=2,\cdots K_{\max}} \{VF(k)\}$ and $K_{opt} = \arg\min_{k=2,\cdots K_{\max}} \{VF(k)\}$;

*Step* 4. Denote $\mathbf{C}_{opt} = \{C_1, C_2, \cdots, C_{K_{opt}}\}$;

*Step* 5. Select the cluster with $k^* = \arg\max_{k=1,2,\cdots \mathrm{card}\{\mathbf{C}\}} \{Intra(k)\}$ from $\mathbf{C}_{opt}$ and denote it by $C_{k^*}$;

*Step* 6. Calculate the optimal $K'_{opt}$ of $C_{k^*}$ and divide $C_{k^*}$ into $K'_{opt}$ groups $\{C'_1, C'_2, \cdots, C'_{K'_{opt}}\}$ using conventional *K*-means clustering;

*Step* 7. Denote $\mathbf{C} = \mathbf{C}_{opt} \bigcup \{C'_1, \cdots, C'_{K'_{opt}}\} - \{C_{k^*}\}$ and calculate the validity function $VF'$ of $\mathbf{C}$;

*Step* 8. If $VF' < VF_{opt}$, set $VF_{opt} = VF'$ and $\mathbf{C}_{opt} = \mathbf{C}$; Else go to step 8;

*Step* 9. If $\mathrm{card}\{\mathbf{C}_{opt}\} < K_{\max}$, go to step 4;

Else stop the algorithm and output $\mathbf{C}_{opt}$ as the optimal clustering result.

**Fig. (1).** The main steps of the dynamic K-means clustering.

### 2.2. The Dynamic *K*-means Clustering Method

As described above, due to the outliers and diversity of samples, severe heterogeneity may exist in one or more groups obtained by conventional *K*-means clustering method. Moreover, this problem could not be solved by simply increasing the value of *K*. To improve the result of conventional *K*-means clustering, a dynamic *K*-means clustering method is developed.

The dynamic clustering is an iterative *K*-means clustering method. The main difference between dynamic clustering and conventional *K*-means clustering is that in dynamic clustering, the number of clusters is not constant. In each iteration of the dynamic clustering, the goodness of the *k*th cluster obtained is evaluated by the index *Intra*(*k* ) defined in formula (2). Since this index represents the compactness of a cluster, it is implied that a cluster with larger *Intra*(*k* ) may consist of samples with severe heterogeneity. Therefore, the cluster with 'worst' (largest) *Intra*(*k* ) is picked out and divided into sub-clusters iteratively. The main steps of the dynamic *K*-means clustering method are shown in Fig. (**1**).

There are two stages in Algorithm 1. In the first stage (step 1 to 4), samples are divided into $K_{opt}$ clusters by conventional *K*-means clustering. The maximum *K* is set to be $\sqrt{N}$ according to Wu and Yang [34]. The second stage (step 5 to 9) is a process of iterative partitioning in order to optimize the validity function. Since the total number of clusters increases in each iteration, the convergence of the algorithm is ensured.

### 2.3. Logistic Regression

Suppose there are *N* observations and each observation $X_i$, ($i$ = 1, 2,..., *N* )can be described as $X_i = [x_{i,1}, x_{i,2}, ..., x_{i,D}, y_i]$, where $x_{i,d}$ ($d$ = 1, 2..., *D*) are observed values of attributes (independent variables), and $y_i \in \{0,1\}$ is the occurrence of state of nature (target variable). The purpose of logistic regression is to investigate the relationship between the independent variables and target variable. Generally, the logistic regression model for *D* independent variables $x_1, x_2, ..., x_D$ can be written as

$$P(y=1) = 1/\{1+\exp[-(\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_D x_D)]\} \quad (6)$$

where $\alpha, \alpha_1, \alpha_2, ..., \alpha_D$ are regression coefficients. In producing the logistic regression equation, the maximum-likelihood method is commonly used to estimate the coefficients and determine the statistical significance of the variables [35].

After the establishment of the logistic regression model, it can be utilized to estimate the probability $P(y_j^p = 1)$ for new-arrived samples $X_1^p, X_2^p, ..., X_M^p$ whose values of independent variables $x_{i,d}^p$ ($j$ = 1,2, ..., *M* and $d$ = 1,2, ..., *D*) are observed but target variable $y_i^p$ is unknown, using the following

$$P(y_j^p = 1) = 1/\{1+\exp[-(\hat{\alpha}_0 + \hat{\alpha}_1 x_{j,1} + \hat{\alpha}_2 x_{j,2} + \cdots + \hat{\alpha}_D x_{j,D})]\} \quad (7)$$

where $\hat{\alpha}_i$ ($i$ = 1, 2,..., *D*) are the estimated coefficients of the logistic regression model.

### 2.4. The Risk Decision-making Approach

After the samples are divided into clusters $C_1, C_2, ..., C_K$, relationship between independent variables and target variable in each cluster are modelled by logistic regression. Consequently, the smuggling probabilities of samples could be obtained. As described in Section 1, in China's customs targeting problem, there are two states of nature, one state is legality and the other is smuggling, which means the set of states of nature is $\Omega = \{\theta_1$ (*legality*), $\theta_2$ (*smuggling* )$\}$. For a newly encountered declaration of import/export goods, the customs officials have to make a decision, and the set of actions is $A = \{a_1$ (*no − inspection*), $a_2$ (*inspection*)$\}$ . Since they do not know exactly whether a declaration of goods is of smuggling, they can make the inspection decision by a decision rule $\delta(p, a)$ : $[p \le p \rightarrow a_1, a_2]$ ($[E \rightarrow u, v]$ means that if *E* is true then take action $a_1$, otherwise take $a_2$ ), where $p = \Pr(\theta_2)$ is the estimated smuggling probability of the declaration of goods and $p$ is a threshold smuggling probability.

This decision rule indicates that, if $p \le p$, the decision-maker predicts that the state of $\theta_1$ would occur, and then he/she takes action $a_1$ ; on the contrary, if $p > p$ , the decision-maker predicts that the state of $\theta_2$ would occur and then takes action $a_2$.

Two critical parameters of the above decision rule is the estimated smuggling probability $p = \Pr(\theta_2)$ of the declaration of goods and the threshold smuggling probability $p$. Reasonably, the newly encountered declared goods should be assigned to a specific cluster first, and then the logistic regression model in the cluster is used to estimate the probability of smuggling. Assume that the newly encountered declaration of goods is assigned to cluster $C_k$ , the empirical risk of inspection decision in cluster $C_k$ is

$$E_k = \sum_{i=1}^{n_k} (sign(p_i^k - p_0^k) - y_i)^2 \tag{8}$$

where $n_k$ is the number of historical observations in cluster $C_k$, $p^k$ is the threshold probability in cluster $C_k$ , and $y_i$ is the occurrence of smuggling ( $y_i = 1$ means smuggling and $y_i = 0$ means not smuggling). To minimize the empirical risk of inspection decision, the threshold probability $p^k$, ($k = 1, 2,..., K$) is determined by

$$p_0^k = \arg\min_{p\in[0,1]} \sum_{i=1}^{n_k} (sign(p_i^k - p) - y_i)^2 \tag{9}$$

To sum up, the main procedure of the risk decision making rule (Decision Making Rule-I, DMR-I) is shown in Fig. (**2**).

Algorithm 2 The risk decision-making rule (DMR-I)

Input: $n$ historical observations $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n$

$m$ observations $\mathbf{X}_1^p, \mathbf{X}_2^p, \cdots, \mathbf{X}_m^p$ with unknown states of smuggling

Output: Inspection decision for each observation $\mathbf{X}_j^p$, $j = 1, 2, \cdots, m.$

*Step* 1. Using Algorithm 1 to divide $n$ historical observations into $K$ clusters ;

*Step* 2. In each cluster $C_k$, $(k = 1, 2, \cdots, K)$, establish a logistic regression model and determine the optimal threshold probability $p_0^k$ using Eq. (9);

*Step* 3. Loop for each observation $\mathbf{X}_j^p$, $(j = 1, 2, \cdots, m)$:

i) assign it to the nearest cluster $C_k$;

ii) use the logistic regression model in $C_k$ to estimate its smuggling probability $p_j^k$;

iii) if $p_j^k > p_0^k$ ..... inspection action on the $\mathbf{X}_j^p$.

**Fig. (2).** The main steps of DMR-I.

In the application to China's customs targeting problem, the risk decision-making approach could be summarized as follows. Historical observations are divided into a number of clusters by using the dynamic $K$-means clustering method, and the relationship between the occurrence probabilities of smuggling and the attributes of observations is modelled by logistic regression in each cluster. After the estimated smuggling probabilities are obtained, a threshold probability is determined such that the empirical error is minimized. Given a newly encountered declaration of goods whose state of smuggling is unknown, it is first assigned to a cluster, and then the probability of smuggling is estimated by the logistic regression model. By comparing the estimated probability of smuggling with the threshold probability in this cluster, the state of smuggling is predicted and consequently, an action is suggested.

Due to the time and resource constraints on decision's execution, declared goods with smuggling probabilities larger than the threshold probability in some clusters may not be totally inspected. As a matter of fact, most customs are only able to inspect a maximum of 5% of declared goods. To make the decision rule applicable in such situations, the DMR-I

is modified and an alternative decision making rule (Decision Making Rule-II, DMR-II) is developed as follows. Denote by max_ per the maximum percentage of goods that the customs official could inspect within their execution time and ability. If in cluster $C_k$, $n_{inspect} = count\left\{ p_i^k \geq p_0^k \mid i = 1,\ldots,n_k \right\} > \max\_\text{per} \bullet n_k$, which means the number of goods suggested to be inspected is larger than the maximum capacity of inspection, the declared goods with highest smuggling probabilities are suggested to be inspected within the capacity of inspection.

## 3. CASE STUDY

In this section, we report the case study and results. The description of data is presented first, followed by variable selection for clustering and logistic regression. To illustrate the characteristics of the samples, we apply conventional *K*-means clustering and the proposed dynamic *K*-means clustering algorithm to them. In the end, we compare the results of different decision-making approach using different clusters algorithms.

### 3.1. Data Description

The data set used in case study is collected from the database of one local customs in China during on year. The number of observations in the data set is 300,825. For each sample indicating a declaration of import/export goods in the data set, there are one label that indicates whether it is of smuggling and more than 30 attributes such as transportation mode of goods, delayed days of declaration, quantity, net weight, declared unit price of goods, *et al*.

For the performance evaluation, cross validation is often used to obtain robust and consistent predictions [36]. In this procedure, a data set is randomly partitioned into two subsets: a training data set and a testing data set. The training data set is used to construct the prediction models, and the testing data set is used to evaluate their performance. In this study, the data set consists of 200,000 observations from the first eight months is used as the training data set, and the remaining 100,825 observations are used as the testing data set. In both training and testing data sets, the smuggling ratio (number of smuggling goods/total number of goods) is about 5%. Since the customs aim to inspect more smuggling goods out of a small proportion of declared goods, the hit ratio is used to evaluate the performance of risk decision-making approach, which is defined as

$$\text{hit ratio} = \frac{\text{number of smuggling goods caught by the decision rule}}{\text{number of goods that are suggested to be inspected}}.$$

Clearly, a higher hit ratio, which means more smuggling declared goods are caught within a given number of inspected goods, indicates a better decision-making result.

### 3.2. Variables Selection in Clustering and Logistic Regression

In order to apply clustering method to the data set for preprocessing, variables (attributes) used for clustering should be first selected. Recall that the aim of clustering is to group the large amount of declared goods into a number of clusters so that samples in each cluster are approximately homogeneous. By discussing with the officials of China's customs, eight variables that represent the similarity of declaration of import/export goods are selected for clustering. These variables are described in Table **1**.

**Table 1. Selected variables for clustering.**

| Variable | Description | Type |
|---|---|---|
| DMG | Declared measure of goods | Categorical |
| LDM | Legal declared measure of goods | Categorical |
| FCO | Goods belong to an easy declaration of false country of origin | Binary |
| UPG | Declared unit price of goods | Continuous |
| RVQ | Ratio of value to quantity | Continuous |
| VNW | Ratio of value to net weight of goods | Continuous |
| QNW | Ratio of quantity to net weight of goods | Continuous |
| GNW | Ratio of gross weight to net weight | Continuous |

For logistic regression that estimates the empirical smuggling probabilities of goods, the target variable is the state of smuggling or not, which means that the variable is binary with the value 1 or 0, and the independent variables are a

number of attributes of goods. Based on custom officials' experience, we select thirteen independent variables indicative of the smuggling probability of goods. These variables are shown in Table **2**.

**Table 2. Selected variables used in logistic regression analysis.**

| Variable | Description | Type |
|---|---|---|
| GNC | Whether import/export of the goods needs specific certificates | Binary |
| FCO | Goods belong to an easy declaration of false country of origin | Binary |
| SRC | Importer's/exporter's classification based on its smuggling records | Categorical |
| IEO | Ownership of the importer/exporter | Categorical |
| MOT | Modes of trade | Categorical |
| TMG | Transportation mode of goods | Categorical |
| WMG | Wrap mode of goods | Categorical |
| DDD | Delayed days of declaration | Continuous |
| QWR | Ratio of quantity to net weight | Continuous |
| VWR | Ratio of value to net weight of goods | Continuous |
| GNR | Ratio of gross weight to net weight | Continuous |
| SSL | Synthesized smuggling likelihoods of importer/exporter and declarer | Continuous |
| UPG | Declared unit price of goods | Continuous |

### 3.3. The Characteristics of Data

Due to the variety and diversity of export/import trades in China, the values of variables selected for clustering have abnormally large ranges. For example, the values of variable VNW (ratio of value to net weight of goods) and QNW (Ratio of quantity to net weight of goods) range from less than one to more than a million. Tables **3** and **4** report the percentage of observations that lie in each interval of $\log_{10}$ (VNW) and $\log_{10}$ (QNW), respectively.

**Table 3. The percentage of observations in each interval of $\log_{10}$ (VNW).**

| Interval | $[-\infty, 1]$ | [1, 2] | [2, 3] | [3, 4] | [4, 5] | [5, 6] | [6, 7] |
|---|---|---|---|---|---|---|---|
| Percentage | 87.38% | 8.61% | 2.96% | 0.88% | 0.12% | 0.044% | 0.007% |

**Table 4. The percentage of observations in each interval of $\log_{10}$ (QNW).**

| Interval | $[-\infty, 1]$ | [1, 2] | [2, 3] | [3, 4] | [4, 5] | [5, 6] | [6, 7] |
|---|---|---|---|---|---|---|---|
| Percentage | 69.56% | 14.20% | 10.34% | 4.42% | 1.39% | 0.084% | 0.014% |

It can be observed from Table **3** and **4** that most values of the variables VNW and QNW lie in the interval (0,100), and a small number of values are more than a million. As pointed out by Xu *et al*. [37], there is dead-unit problem in the method of conventional *K*-mean clustering. That is, if some units are initialized far away from the input data set in comparison with other units, they then immediately become dead without learning chance any more in the whole learning process. Therefore, in *K*-means clustering, those samples with abnormally large variable values would make a large number of samples with relatively small or medium values clustered into one large group. Consequently, statistical analysis would be hindered due to the heterogeneity among the samples.

In order to investigate the impact of the abnormally large variable values on the clustering results, conventional *K*-means clustering method is used to divide the data set into clusters. To determine the optimal cluster number, we first calculate the validity function values for different *K*. As suggested in Wu and Yang [34], the range of *K* is set to vary from 2 to $\sqrt{N}$ (*N* is the total number of samples). The values of $VF(K)$ corresponding to each *K* are shown in Fig. (**3**) (we omit the values of $VF(K)$ for $K>20$ because those values are all larger than 0.10).

It can be observed from Fig. (**3**) that $K=4$ is the optimal cluster number with the smallest function value 0.082. The result of *K*-means clustering with $K=4$ is shown in Table **5**.

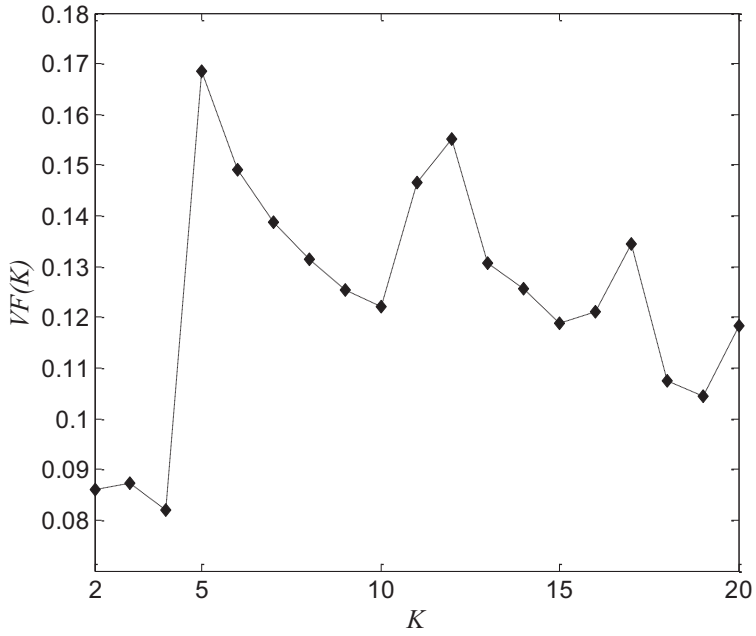


**Fig. (3).** *VF*(*K*) of conventional *K*-means clustering.

**Table 5. The results of conventional *K*-means clustering on the data set.**

| Cluster index | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| Number of samples | 1879 | 3184 | 174647 | 20290 |
| Percentage of samples | 0.94% | 1.59% | 87.32% | 10.15% |

Table **5** illustrates that more than 85 percent of the samples are clustered into one group, which means that there may be severe heterogeneity in this group. In the application of the *K*-means clustering method, it has been found that as the *K* turned larger, number of samples in the abnormally large group did not change significantly, while many clusters became very small with only a few samples, which would make statistical analysis insufficient because of the lack of samples. For example, parts of the clustering results of *K*=20 and *K*=50 are shown in Table **6**.

**Table 6. Abnormality in clustering results (*K*=20 and *K*=50).**

| *K* | No. of samples in the largest group | No. of samples in the smallest group |
|---|---|---|
| 20 | 163089 | 1 |
| 50 | 132174 | 1 |

It can be observed from Table **6** that increasing *K* is not an efficient way to improve the clustering result significantly. In order to get relatively equal-size groups in each of which the samples are approximately homogeneous, a more sophisticated clustering method for data preprocessing is needed in China's customs targeting problem.

**Table 7. The result of the dynamic *K*-means clustering method on the training dataset.**

| Cluster index | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| No. of samples | 1879 | 3184 | 1894 | 20290 | 234 | 4 | 23974 |
| Percentage | 0.94% | 1.59% | 0.95% | 10.15% | 0.12% | 0.002% | 11.99% |
| **Cluster index** | **8** | **9** | **10** | **11** | **12** | **13** | |
| No. of samples | 40623 | 6 | 17 | 54622 | 14435 | 38838 | |
| Percentage | 20.31% | 0.003% | 0.009% | 27.31% | 7.22% | 19.42% | |

Alternatively, we apply the proposed dynamic *K*-means clustering method to the training data set. The values of validity function corresponding to different *K* are shown in Fig. (**4**).

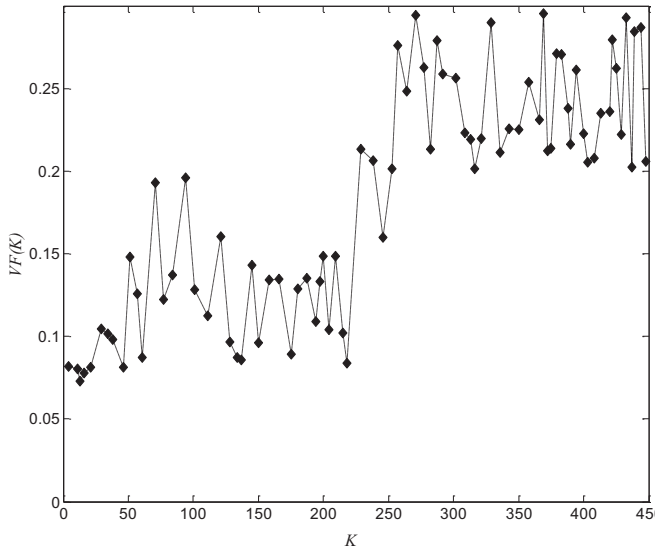


**Fig. (4).** *VF*(*K*) of the dynamic *K*-means clustering method.

It can be observed from Fig. (**4**) that *VF* (*K* ) is minimized when *K*=13, which indicates the clustering result obtained in the iteration corresponding to *K*=13 is optimal. The clustering result is summarized in Table **7**.

It can be observed from Table **7** that the data set could be divided into more clusters with approximately equal sizes by the proposed dynamic *K*-means clustering method. In order to investigate the validity of results obtained by the dynamic *K*-means clustering, we compare the values of validity function for $K \in [8, 20]$ in conventional *K*-means clustering and dynamic *K*-means clustering, as shown in Fig. (**5**).

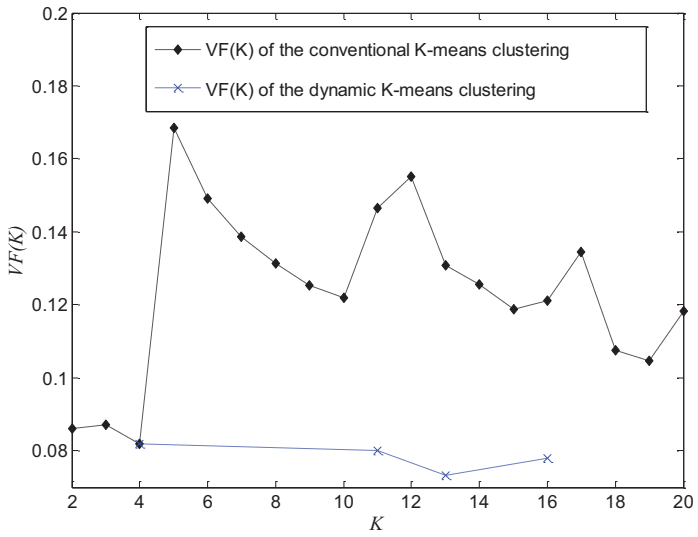


**Fig. (5).** Comparison of *VF*(*K*) in different clustering methods.

By comparing *VF* (*K*) of conventional *K*-means clustering to that of the dynamic *K*-means clustering shown in Fig. (**5**), it can be concluded that for the same cluster number, the dynamic *K*-means clustering method delivers better clustering result with smaller value of validity function. Therefore, the clusters obtained are more compact and separated from each other. Consequently, the heterogeneity may be reduced in these clusters.

### 3.4. Application of the Proposed Risk Decision-making Approach

In order to evaluate the performance of the proposed risk decision-making approach for customs targeting, three different experiments are conducted for comparisons. In the first experiment, logistic regression is directly applied to the calibration data set and validation data set without any preprocessing. In the second experiment, DMR-I and DMR-II are applied, but the conventional *K*-means clustering instead of the dynamic *K*-means clustering is utilized for preprocessing. In the third experiment, DMR-I and DMR-II are applied.

By using SAS System for Windows 9.0 to execute the clustering methods and logistic regression, the computational results of the three experiments are shown in Table **8**. It is noteworthy that the robustness of the k-means clustering algorithm is affected by the initial set of centroids as well as the number of iterations [38]. However, since we focus on developing a dynamic clustering algorithm suitable for partitioning large amount of data and improving the accuracy of risk decision-making, we ignore the issues of initial set of centroids and the number of iterations, simply adopting the default setting in the SAS system.

**Table 8. The results of three different experiments.**

| Decision Rule | Hit Ratio | | |
|---|---|---|---|
| | The 1st experiment | The 2nd experiment | The 3rd experiment |
| DMR-I | 5.83% | 6.02% | 7.34% |
| DMR-II* | 7.18% | 7.49% | 9.81% |

*In DMR-II, max_per is set as 5% in accordance with the maximum inspection capacity of most customs in China.

It can be observed from Table **8** that the accuracy of inspection decision in the 3rd experiment is the best, which indicates that the dynamic *K*-means clustering method is effective in preprocessing data sets that contain a large variety of observations with severe heterogeneity. This is favourable for customs administration, whose purpose is to inspect as more smuggling goods as possible given a limited inspection capacity (about 5% of the total import/export goods).

## CONCLUSION

In this paper, a dynamic *K*-means clustering method has been developed and used as the preprocessing technique for statistical analysis in China's customs targeting problem. The dynamic *K*-means clustering is an iterative *K*-means clustering method whose purpose is to reduce the heterogeneity existing in the samples and consequently, improve the performance of clustering. Compared with conventional *K*-means clustering, the number of clusters varies according to the characteristics of the current clustering result in dynamic *K*-means clustering. Consequently, the data set is divided into approximately equal-size clusters, each of which is more compact and separated from others. Based on the dynamic *K*-means clustering method and logistic regression, a risk decision making approach is proposed. Application results to China's customs targeting problem indicates that the proposed risk decision making approach is effective in improving the accuracy of customs targeting.

The limitations and weakness of our study are that it does not take the accumulation of new-arrived data into consideration and update the decision-making model, and the feature selection for both clustering and logistic regression are arbitrary. Therefore, this study can be further extended along the following two lines. First, as more and more samples (declarations of import/export good) are accumulated as a data stream, the ability of incremental and online learning has become important for accurate and efficient decision-making. Second, optimal set of attributes for logistic regression may be different in the obtained clusters. It is promising to refine the set of attributes used for logistic regression based on the characteristics of samples in each cluster.

## CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

## ACKNOWLEDGEMENTS

We are grateful to the editor and the anonymous reviewers for their helpful and constructive comments and suggestions, which have significantly improved the paper.

This research was supported by the National Natural Science Foundation of China (Grants No. 71471022).

## REFERENCES


[1] B.J. Cohen, "Is expected utility theory normative for medical decision making?", *Med. Decis. Making,* vol. 16, no. 1, pp. 1-6, 1996. [http://dx.doi.org/10.1177/0272989X9601600101] [PMID: 8717589]

[2] L. Ekenberg, M. Danielson, and M. Boman, "Imposing security constraints on agent-based decision support", *Decis. Support Syst.,* vol. 20, no. 1, pp. 3-15, 1997. [http://dx.doi.org/10.1016/S0167-9236(96)00072-3]

[3] F.C. Chu, and J.Y. Halpern, "Great expectations. Part II: generalized expected utility as a universal decision rule", *Artif. Intell.,* vol. 159, no. 1-2, pp. 207-229, 2004. [http://dx.doi.org/10.1016/j.artint.2004.05.007]

[4] T.S. Lee, C.C. Chiu, Y.C. Chou, and C.J. Lu, "Mining the customer credit using classification and regression tree and multivariate adaptive regression splines", *Comput. Stat. Data Anal.,* vol. 50, no. 4, pp. 1113-1130, 2006. [http://dx.doi.org/10.1016/j.csda.2004.11.006]

[5] D. Martens, B. Baesens, T.V. Gestel, and J. Vanthienen, "Comprehensible credit scoring models using rule extraction from support vector machines", *Eur. J. Oper. Res.,* vol. 183, no. 3, pp. 1466-1476, 2007. [http://dx.doi.org/10.1016/j.ejor.2006.04.051]

[6] M.K. Lim, and S.Y. Sohn, "Cluster-based dynamic scoring model", *Expert Syst. Appl.,* vol. 32, no. 2, pp. 427-431, 2007. [http://dx.doi.org/10.1016/j.eswa.2005.12.006]

[7] L. Wang, J. Chu, and J. Wu, "Selection of optimum maintenance strategies based on a fuzzy analytic hierarchy process", *Int. J. Prod. Econ.,* vol. 107, no. 1, pp. 151-163, 2007. [http://dx.doi.org/10.1016/j.ijpe.2006.08.005]

[8] C.Y. Peng, B.D. Manz, and J. Keck, "Modeling categorical variables by logistic regression", *Am. J. Health Behav.,* vol. 25, no. 3, pp. 278-284, 2001. [http://dx.doi.org/10.5993/AJHB.25.3.15] [PMID: 11322627]

[9] N.H. Mayer, "M. Rosenfeld, J. Emerson, C.H. Goss, M.L. Aitken, “Developing cystic fibrosis lung transplant referral criteria using predictors of two year survival", *Am. J. Respir. Crit. Care Med.,* vol. 166, pp. 1550-1555, 2002. [http://dx.doi.org/10.1164/rccm.200202-087OC] [PMID: 12406843]

[10] Z.S. Hua, and B. Zhang, "A hybrid support vector machines and logistic regression approach for forecasting intermittent demand of spare parts", *Appl. Math. Comput.,* vol. 181, no. 2, pp. 1035-1048, 2006. [http://dx.doi.org/10.1016/j.amc.2006.01.064]

[11] Z.S. Hua, B. Zhang, J. Yang, and D.S. Tan, "A new approach of forecasting intermittent demand for spare parts inventories in the process industries", *J. Oper. Res. Soc.,* vol. 58, no. 1, pp. 52-61, 2007. [http://dx.doi.org/10.1057/palgrave.jors.2602119]

[12] Y. Liu, Z.P. Fan, and Y. Zhang, "Risk decision analysis in emergency response: a method based on cumulative prospect theory", *Comput. Oper. Res.,* vol. 42, pp. 75-82, 2014. [http://dx.doi.org/10.1016/j.cor.2012.08.008]

[13] Y. Liu, Z.P. Fan, Y. Yuan, and H. Li, "A FTA-based method for risk decision-making in emergency response", *Comput. Oper. Res.,* vol. 42, pp. 49-57, 2014. [http://dx.doi.org/10.1016/j.cor.2012.08.015]

[14] Z.S. Hua, S.J. Li, and Z. Tao, "A rule-based risk decision making approach and its application in China’s Customs inspection decision", *J. Oper. Res. Soc.,* vol. 57, no. 11, pp. 1313-1322, 2006. [http://dx.doi.org/10.1057/palgrave.jors.2602142]

[15] J. Han, and M. Kamber, *Data mining: Concepts and Techniques.,* Morgan Kaufmann: New York, 2006.

[16] K.A. Smith, R.J. Willis, and M. Brooks, "An analysis of customer retention and insurance claim patterns using data mining: a case study", *J. Oper. Res. Soc.,* vol. 51, pp. 532-541, 2000. [http://dx.doi.org/10.1057/palgrave.jors.2600941]

[17] A. Sfetsos, "Short-term load forecasting with a hybrid clustering algorithm", *IEE Proc-Generation, Transmis. Distrib.,* vol. 150, no. 3, pp. 257-262, 2003. [http://dx.doi.org/10.1049/ip-gtd:20030200]

[18] N.G. Pavlidis, D.K. Tasoulis, and M.N. Vrahatis, "Financial forecasting through unsupervised clustering and evolutionary trained neural networks", *Congr. Evol. Comput.,* vol. 4, pp. 2314-2321, 2003.

[19] M.S. Bhargavi, and S.D. Gowda, "A novel validity index with dynamic cut-off for determining true clusters", *Patt. Recog.,* vol. 48, pp. 3673-3687, 2015. [http://dx.doi.org/10.1016/j.patcog.2015.04.023]

[20] I. Benítez, A. Quijano, J.L. Díez, and I. Delgado, "Dynamic clustering segmentation applied to load profiles of energy consumption from Spanish customers", *Electr. Power Ener. Syst.,* vol. 55, pp. 437-448, 2014. [http://dx.doi.org/10.1016/j.ijepes.2013.09.022]

[21] C. Ozturk, E. Hancer, and D. Karaboga, "Dynamic clustering with improved binary artificial bee colony algorithm", *Appl. Soft Comput.,* vol. 28, pp. 69-80, 2015.
[http://dx.doi.org/10.1016/j.asoc.2014.11.040]

[22] S.L. Des Jardins, "Assessing the effects of changing institutional aid policy", *Res. Higher Educ.,* vol. 42, no. 6, pp. 653-678, 2001.
[http://dx.doi.org/10.1023/A:1012249427051]

[23] D. Lowe, and J. Parvar, "A logistic regression approach to modeling the contractor's decision to bid", *Construct. Manag. Econ.,* vol. 22, no. 6, pp. 643-653, 2004.
[http://dx.doi.org/10.1080/0144619031000164905 6]

[24] S. Pourahmad, S.M. Ayatollahi, S.M. Taheri, and Z.H. Agahi, "Fuzzy logistic regression based on the least squares approach with application in clinical studies", *Comput. Math. Appl.,* vol. 62, no. 9, pp. 3353-3365, 2011.
[http://dx.doi.org/10.1016/j.camwa.2011.08.050]

[25] V. Lacagnina, M.S. Leto-Barone, S. La Piana, G. La Porta, G. Pingitore, and G.D. Lorenzo, "Comparison between statistical and fuzzy approaches for improving diagnostic decision making in patients with chronic nasal symptoms", *Fuzzy Sets Syst.,* vol. 237, no. 16, pp. 136-150, 2014.
[http://dx.doi.org/10.1016/j.fss.2013.10.013]

[26] M.N. Vrahatis, B. Boutsinas, P. Alevizos, and G. Pavlides, "The new k-windows algorithm for improving the k-mean clustering algorithm", *J. Complexity,* vol. 18, no. 1, pp. 375-391, 2002.
[http://dx.doi.org/10.1006/jcom.2001.0633]

[27] J.C. Bezdek, "Cluster validity with fuzzy sets", *J. Cybern.,* vol. 3, pp. 58-73, 1974.
[http://dx.doi.org/10.1080/01969727308546047]

[28] X.L. Xie, and G. Beni, "A validity measure for fuzzy clustering", *IEEE T. Patt. Anal. Mach. Intell.,* vol. 13, no. 8, pp. 841-847, 1991.
[http://dx.doi.org/10.1109/34.85677]

[29] R.N. Dave, "Validating fuzzy partition obtained through c-shell clustering", *Pattern Recognit. Lett.,* vol. 17, no. 6, pp. 613-623, 1996.
[http://dx.doi.org/10.1016/0167-8655(96)00026-8]

[30] J.C. Bezdek, and N.R. Pal, "Some new indexes of cluster validity", *IEEE T. Syst. Man Cybern. -Part B: Cybern.,* vol. 28, no. 3, pp. 301-315, 1998.
[http://dx.doi.org/10.1109/3477.678624]

[31] N. Zahid, M. Limouri, and A. Essaid, "A new cluster-validity for fuzzy clustering", *Patt. Recog.,* vol. 32, no. 7, pp. 1089-1097, 1999.
[http://dx.doi.org/10.1016/S0031-3203(98)00157-5]

[32] S. Bandyopadhyay, and U. Maulik, "Nonparamatric genetic clustering: comparison of validity indices", *IEEE T. Syst., Man Cybern. -Part C. Appl. Rev.,* vol. 31, no. 1, pp. 120-125, 2001.

[33] K.R. Zalik, and B. Zalik, "Validity index for clusters of different sizes and densities", *Pattern Recognit. Lett.,* vol. 32, no. 2, pp. 221-234, 2011.
[http://dx.doi.org/10.1016/j.patrec.2010.08.007]

[34] K.L. Wu, and M.S. Yang, "A cluster validity index for fuzzy clustering", *Pattern Recognit. Lett.,* vol. 26, no. 9, pp. 1275-1291, 2005.
[http://dx.doi.org/10.1016/j.patrec.2004.11.022]

[35] D.W. Hosmer, and S. Lemeshow, *Applied Logistic Regression.,* John Wiley & Sons: New York, 2000.
[http://dx.doi.org/10.1002/0471722146]

[36] K. Coussement, and W. Buckinx, "A probability-mapping algorithm for calibrating the posterior probabilities: A direct marketing application", *Eur. J. Oper. Res.,* vol. 214, no. 3, pp. 732-738, 2011.
[http://dx.doi.org/10.1016/j.ejor.2011.05.027]

[37] L. Xu, A. Krzyzak, and E. Oja, "Rival penalized competitive learning for clustering analysis, RBF net, and curve detection", *IEEE Trans. Neural Netw.,* vol. 4, no. 4, pp. 636-649, 1993.
[http://dx.doi.org/10.1109/72.238318] [PMID: 18267764]

[38] F. Lolli, A. Ishazaka, and R. Gamberini, "New AHP-based approaches for multi-criteria inventory classification", *Int. J. Prod. Econ.,* vol. 156, pp. 62-74, 2014.
[http://dx.doi.org/10.1016/j.ijpe.2014.05.015]