Association for Information Systems
AIS Electronic Library (AISeL)

PACIS 2014 Proceedings

Pacific Asia Conference on Information Systems (PACIS)

2014

AN EFFICIENT MULTI-CRITERIA DECISION-MAKING APPROACH BASED ON HYBRIDIZING DATA MINING TECHNIQUES AN EFFICIENT MULTI-CRITERIA DECISION-MAKING APPROACH BASED ON HYBRIDIZING DATA MINING TECHNIQUES

Kuang Yu Huang
Ling Tung University, kyhuang@teamail.ltu.edu.tw

Yu Chun Huang
National Chiao Tung University, may158182@gmail.com

AN EFFICIENT MULTI-CRITERIA DECISION-MAKING APPROACH BASED ON HYBRIDIZING DATA MINING TECHNIQUES

Kuang Yu Huang, Department of Information Management, Ling Tung University, Taichung, Taiwan, R.O.C., kyhuang@teamail.ltu.edu.tw

Yu Chun Huang, Department of Information Management and Finance, National Chiao Tung University, Hsinchu, Taiwan, R.O.C., may158182@gmail.com

Abstract

Multiple-criteria decision-making (MCDM) that deals with multiple criteria in decision-making environments has been explicitly applied to various decision-making fields. Nevertheless, the critical issues of uncertainty and inaccuracy generally and gradually exists in the majority of the MCDM processes because of (1) prejudice and preference of decision-makers or experts as well as (2) the insufficiency information of the input and output. Therefore, this research efficiently proposed a novel method, FVM-index method, to resolve the limitations happened when MCDM is applied. The FVM-index approach, which consists of the fuzzy set theory (FST), the variable precision rough set (VPRS), and the cluster validity index (CVI) function, not only provides optimized classification results for the datasets but also filters out the uncertainty and inaccuracy instances from surveyed datasets by VPRS theory. Because the datasets are refined by the proposed FVM-index method, the decision makers will be able to effectively obtain the suitable results of MCDM.

Keyword: Multi-Criteria Decision Making, Variable Precision Rough Set, Fuzzy Set Theory, FVM-index method.

1 INTRODUCTION

Decision making is extremely intuitive when considering single criterion problem, since we only need to choose the alternative with highest preference rating. But nowadays, the decision issues may no longer simply be solved by decision-making methods for the single destination but handled by multi-objective decision-making methods. When decision makers evaluate alternatives with multiple criteria, many problems, such as weights of criteria, preference decencies, and conflicts among criteria, seem to complicate the problems and need to be overcome by sophisticated methods.

Therefore, we will discuss about multi-criteria decision-making that is a system with multiple conditional and multi-decision attributes in this research. Multi-criteria decision-making, MCDM, which also be called MCDA, is a study which has made apparent progress in the past two decades.

Lioua and Tzeng (2012), and Tzeng and Huang (2011) both illustrate the primary steps of Multi-criteria decision-making (MCDM) methods. These steps can be divided into three stages: (1) Data Processing / Statistical and Multivariate Analysis (2) Planning / Designing (3) Evaluating / Choosing. This study proposes a hybrid data-mining approach that is applied to Data Processing / Statistical and Multivariate Analysis aspect to help building an efficient and effective MCDM system.

When it comes to MCDA, nowadays, the majority of the researchers applied the MCDA's skills to the final Evaluating / Choosing stage. The first restriction that these techniques applied to the final Evaluating / Choosing stage would face is that the preferences of decision makers and experts' advices have to be considered as well. Moreover, the relations (Conditions on conditions, conditions on decision-making, and decision making on decision making) between attributes are extremely complicated for a multi-input (conditional attributes and independent variables) and multi-output (decision attributes and dependent variables) information system.

Therefore, this research intends to propose a hybrid data-mining technique to eliminate the uncertain instances from the original datasets in the Data Processing / Statistical and Multivariate Analysis stage. Then, the decision-makers could apply these extrated-accurate instances to derivate the reliable decision-making rules in third Evaluating / Choosing stage.

The remainder of this paper is organized as follows. In Section Two, we present the fundamental principles of the Fuzzy Set theory, the VPRS theory, and the proposed CVI function. In Section Three, we describe the integration of these concepts to create the proposed FVM-index method. In Section Four, we illustrate the classification results of the proposed method when applied to a hypothetical dataset. Finally, In Section Five, we present some brief concluding remarks and indicates the intended direction of future research.

2 REVIEW OF RELATED METHODOLOGIES

2.1 Index Function (Huang 2009)

Assume that each object x_i in the dataset has just m attributes and the l-th attribute a_l can be divided into p_l clusters, then $C_{a_l}(x_i)$ gives the index of the cluster to which the l-th attribute a_l of object x_i belongs. Here $C_{a_l}(x_i)$ is given by $C_l(x_i) = I_{\max}(\mu_j(x_i(a_l))) =$ Index $(\max(\mu_j(x_i(a_l))))$ *for* $1 \le l \le m, 1 \le i \le n$, where $I_{\max}(\mu_j(x_i(a_l)))$ returns the index of the cluster corresponding to the maximum value of the membership functions of the l-th attribute of x_i.

2.2 The VPRS Model Index Function (Huang 2009)

The basic principles and notations of information systems (S) and the application of VPRS theory (Ziarko 1993) to the processing of such systems are described in the following sections.

2.2.1 β -lower and β -upper approximate sets

A typical information system has the form $S = (U, A, V_q, f_q)$, where U is a non-empty finite set of records, A is a non-empty finite set of attributes describing these records, $X \subseteq U$, and $R \subseteq A$. Generally speaking, the attributes in set A can be partitioned into a set of conditional attributes $C \neq \phi$ and another set of decision attributes $D \neq \phi$, i.e. $A = C \bigcup D$ and $C \bigcap D = \phi$. For each attribute, $q \in A$, V_q represents the domain of q, i.e. $V = \bigcup V_q$. Finally, $f_q : U \times A \to V$ is an information function defined such that $f(x, q) \in V_q$ for $\forall q \in A$ and $\forall x \in U$.

The VPRS method used in this study applies the systematic method presented by the current author (Huang 2009) to decide a suitable value of the threshold parameter β, i.e., the value of β at which a certain proportion of the records in a specific conditional class are classified into the same decision class. When processing an information system using a VPRS model with $0.5 < \beta \leq 1$, the objective is to recognize the β -lower and β -upper approximate sets in terms of each class of the decision attribute. In general, the β -lower approximation sets $X \subseteq U$ and $P \subseteq C$ are given by

$$ {}_{\beta}\underline{R}_P(X) = \{x \in U : P(X/[x]_P) \geq \beta\} = \bigcup \{[x]_P : P(X/[x]_P) \geq \beta\}. $$

Similarly, the β -upper approximation sets $X \subseteq U$ and $P \subseteq C$ can be expressed as

$$ {}_{\beta}\overline{R}_P(X) = \{x \in U : P(X/[x]_P) > 1-\beta\} = \bigcup \{[x]_P : P(X/[x]_P) > 1-\beta\}. $$

Note that $P(X/Y) = |X \bigcap Y| / |Y|$ if $|Y| > 0$, and $P(X/Y) = 1$ otherwise. Note also that $|X|$ indicates the cardinality of set X.

Ziarko (1993) defined the following alternative expression for the β -boundary region of X in S:

$$ {}_{\beta}BND_P(X) = \{x \in U : 1-\beta < P(X/[x]_R) < \beta\} = \bigcup \{[x]_P : 1-\beta < P(X/[x]_R) < \beta\}. $$

The accuracy of the VPRS classification results can be quantified as follows:

${}_{\beta}\alpha_c = |{}_{\beta}\underline{R}_P(X)| / |{}_{\beta}\overline{R}_P(X)|$, where $X = \{x : C_d(x) = c, \forall x \in U\}$; and $|{}_{\beta}\underline{R}_P(X)|$ and $|{}_{\beta}\overline{R}_P(X)|$ are the cardinalities of β -lower and β -upper approximate sets, respectively, when ranking the instances (x) in the dataset in terms of the c -th indiscernible discretized Decision Attributes Clusters Vector (DACV).

2.3 The Extension Principle of Fuzzy Sets

2.3.1 Fuzzy C-Means (FCM) method

The fuzzy C-means (FCM) clustering method, developed by Dunn in 1973 and later refined by Bezdek in 1981, has many applications, ranging from feature analysis to clustering and classifier design. The FCM clustering method consists of two basic procedures, namely (i) calculating the cluster centroids within the dataset, and (ii) determining the cluster

memberships of each data object. This two-step procedure is repeated iteratively until the centroids of all the clusters within the dataset converge.

2.3.2 Fuzzy Sets Operator (Tsoukalas & Uhrig 1997; Zadeh 1965)

For a MCDM information system, fuzzy-based clustering methods yield more reliable decision-making rules compared to conventional crisp clustering schemes. In fuzzy sets, all the properties can be expressed using the membership function of the sets involved and the definitions of union, intersection, and complement. For one, if every instance in the dataset has M decision attributes, the j-th and the k-th attribute d_j, and d_k can be clustered into p_j and p_k cluster. Respectively, $\mu(C_{d_j}(x_i))$ and $\mu(C_{d_k}(x_i))$ give the membership functions of the index in the cluster to which the j-th and the k-th decision attributes of instance x_i belonging to. In the FVM-index method proposed in this study, the minimize operator is used to aggregate the membership function values of M decision attributes. The following minimize operator should be applied:

$$\mu_{\min}(C_d(x_i)) = \min(\mu(C_{d_1}(x_i)), \mu(C_{d_2}(x_i)), \ldots, \mu(C_{d_M}(x_i)))$$

$$= \mu(C_{d_1}(x_i)) \wedge \mu(C_{d_2}(x_i)) \wedge \ldots \wedge \mu(C_{d_M}(x_i))$$

2.4 The MD index function

Before applying VPRS model to an information system, it necessitates the number of clusters for each attribute in the dataset. Unfortunately, such information is not known a priori. Finding the optimal number of clusters for discretizing a set of real-valued attributes is an NP-hard problem (Roy & Pal 2003). To prevent from falling into the NP-hard situation, cluster validity index (Bezdek 1974) is integrated with VPRS model for the assessment of cluster quality in this paper. Namely, we choose the adequate cluster number by means of comparing distinct index values. Many cluster validity indexes have been proposed to assess the nature of fuzzy clustering methods (Ghosh 2011; Pakhira 2004; Wang & Zhang et al. 2007). The common insufficient part of them is that they do not take the complicated interrelationships among various attributes into account. Therefore, the proposed MD index function that modified from the Huang-index (Huang 2010) has the form:

$MD(C) = \left(\frac{1}{C} \times \frac{E_1}{\overline{F}_{N_I}} \times \overline{D}_{N_I} \right)$, where C and $\mathrm{N_I}$ denote the number of clusters of decision attributes and the number of indiscernible DACV respectively, and ${}_{\beta}\alpha_c$ is the classification accuracy when evaluated in terms of the c-th indiscernible DACV. In addition, E_1 is constant, which is equal to $\overline{E_1}$ in the PBMF-index function and $\overline{F}_{N_I}$ is obtained by accumulating the value of E_i' for each indiscernible DACV, where E_i' is given by $E_i' = \sum_{j=1}^{n} {}_d\overline{\mu}_{ji}^{m'} \left\| x_j - z_i' \right\| \Big/ {}_{\beta}\alpha_i$ in which ${}_d\overline{\mu}_{ji}(x_j(d))$ is the aggregated membership function value. The minimize fuzzy set operation is used to process the aggregated membership function value of instance x_j in the clusters indicated by the c-th indiscernible DACV, and z_c' is the multi-dimensional centroid of the β-lower approximate sets associated with the clusters indicated by the c-th indiscernible DACV, and is obtained by computing the mean values of each of the conditional and decision attributes of instances within the corresponding sets. Furthermore, m' is the fuzzification parameter and n is the total number of instances in the dataset. Finally, the value of D'_{N_I} is equal to the maximum separation distance amongst the centroids of the all lower approximate sets associated with each indiscernible

DACV, i.e., $\overline{D}_{N_I} = \max_{i,j=1}^{N_I} \left\| z_i' - z_j' \right\|$.

3 THE PROPOSED FVM-INDEX METHOD

The proposed FVM-index approach not only provides optimized classification results for the datasets but also filters out the uncertainty and inaccuracy instances from surveyed datasets by VPRS theory.

3.1 The index-based classification method for real-valued type of multi-decision attributes

This research integrates FCM method, Fuzzy arithmetic relations, variable precision rough set, and cluster validity index function to get the optimal solution of cluster-classified for multi-decision attributes of continuous value. The detailed steps are presented as follows:

(1) Specify number of clusters per conditional and decision attribute in interval $[2, \mathrm{N}_{\max}]$.

(2) Fuzzify attribute values of instances using FCM method.

(3) Assign each attribute of each instance to appropriate conditional or decision attribute cluster.

Utilize the index function $C_{a_l}(x_i) = I_{\max}(\mu_j(x_i(a_l))) = Index(\max(\mu_j(x_i)))$ for $1 \le l \le m, 1 \le i \le n$, the membership function values of each attribute in each instance are processed in order to determine the conditional or decision attribute cluster to which each attribute belongs and to obtain the corresponding DACV. Then, compute aggregated membership function value and classification accuracy.

Following the steps mentioned above, we can get the decision-making attribute clustering indexes of every attribute that are obtained each real-valued instances. As we obtain DACV, using the Fuzzy arithmetic relations to consolidate every membership function values of this vector into a single membership function value.

(5) Calculate centroids of β-lower approximate sets associated with each indiscernible DACV.

Furthermore, according to the theory of approximation set in section "The MD index function", the β-lower approximation, the β-upper approximation, and the boundary sets of the c-th cluster vector in DACV are obtained. Therefore, by calculating the cardinality ratio of the β-lower approximation set to the β-upper approximation set regarding multi-attribute decision-making datasets, the VPRS classification accuracy in the c-th clustering vector of DACV is obtained.

Assume that there is more than one real-valued attribute in the information systems. The centroids of the β-approximation sets under each DACV have to be calculated the average of each attribute of all instances under the set. (Including conditional and decision attributes.)

(6) Determine value of MD cluster validity index.

Having determined the aggregated membership function values, classificatory accuracy, and centroids of the β-lower approximate sets, the clustering results is evaluated by using the MD index function.

(7) Check termination criterion.

When finishing calculating the cluster validity index value under the specific cluster

number N, check whether the cluster number N exceeds the maximum of the cluster number. If the cluster number N hasn't exceeded the maximum of the cluster number yet, adding one and returning to step one to restart the calculation of the FCM method, Fuzzy arithmetic relations, variable precision rough set (VPRS), and cluster validity index. Through carrying out these procedures, not to end up the first step of the ascending process until the termination criterion has been satisfied, and then moving to the ultimate step.

(8) Identify value of MD cluster validity index.

Once the termination criterion has been satisfied, the values of the MD index function obtained for $N = N_{\min} \sim \mathrm{N}_{\max}$ are compared in order to identify the clustering solution which yields the maximum index function value, i.e., the clustering solution which optimizes both the number of clusters per attribute and the overall classification accuracy of the dataset.

3.2 A step-by-step example showing calculation of MD index value

This section illustrates the derivation of the MD index value for a simple hypothetical dataset comprising just four entries. An assumption is made that each entry has two conditional attributes, a_1, a_2, and two decision attributes, d_1, d_2. Let the four instances be defined as $x_1(-1.40,1.30,1.05,-0.95)$, $x_2(-1.60,1.20,0.95,-0.85)$, $x_3(-1.95,1.45,0.60,-0.70)$ and $x_4(-2.05,1.55,0.50,-0.60)$. In accordance with the FVM-index method, the real-valued instances in the hypothetical dataset are discretized using the FCM technique. Note that an assumption is made that each conditional and decision attribute is partitioned into 2 clusters. The membership function values of each attribute of each instance are summarized in Table 1(a). The attribute values of each instance are then assigned to appropriate conditional or decision attribute clusters by applying the index function $I_{\max}$ to the corresponding membership function values. The mapping results are shown in Table 1(b). As shown, the discretized vectors of the four instances x_i $(I_{a_1}, I_{a_2}, I_{d_1}, I_{d_2})$ have the form $x_1(1,2,1,2)$, $x_2(1,2,1,2)$, $x_3(2,1,2,1)$, and $x_4(2,1,2,1)$, respectively. The β-upper and β-lower approximate sets associated with each indiscernible DACV are calculated in accordance with the formulation given in Section 2.2.3 and are also shown in Table 1(b).

Code of instances	Conditional attributes				Decision attribute			
	a_1		a_2		d_1		d_2	
1	0.975	0.025	0.061	0.939	0.990	0.010	0.025	0.975
2	0.937	0.063	0.025	0.975	0.985	0.015	0.061	0.939
3	0.012	0.988	0.939	0.061	0.015	0.985	0.939	0.061
4	0.008	0.992	0.975	0.025	0.010	0.990	0.975	0.025

Table 1(a) Membership function values of each attribute of each instance

Code of instances	β -lower approximate sets ${}_{\beta}\underline{R}(X : C_D(\mathrm{x}) = c, x \in X)$				
	a_1	a_2	d_1	d_2	
1	1	2	1	2	* ${}_{\beta}\underline{R}(X : C_D(\mathrm{x}) = 1, x \in X)$
2	1	2	1	2	
3	2	1	2	1	* ${}_{\beta}\underline{R}(X : C_D(\mathrm{x}) = 2, x \in X)$
4	2	1	2	1	
* The first and second indiscernible DACVs are [1,2] and [2,1], respectively. # Each of the lower approximate sets ${}_{\beta}\underline{R}(X : C_D(\mathrm{x}) = c, x \in X)$ is equal to the corresponding upper approximate set ${}_{\beta}\overline{R}(X : C_D(x) = c, x \in X)$, ${}_{\beta}\alpha_1 = {}_{\beta}\alpha_2 = 1$.					

Table 1(b) β -lower approximate sets and β -upper approximate sets associated with c -th DACV

Having computed the β -upper and β -lower approximate sets, the FVM procedure then calculates the aggregative membership function values of each instance and the classification accuracy of the clustering solution. Taking the first instance x_1 as an example, the two indiscernible DACVs have values of $C_D(x_1)$= [1,2] and [2,1], respectively (see Table 1(b)). The aggregative membership function value of x_1 in the first DACV [1,2], i.e., ${}_{d}\overline{\mu}_{11}$, is obtained using the minimize operator as min(0.990, 0.975) = 0.975. Similarly, the aggregative membership function value of x_1 in the second DACV [2,1], i.e., ${}_{d}\overline{\mu}_{21}$, is obtained as 0.01. The aggregative membership function values of the four instances in the hypothetical dataset are shown in Table 1(c). The classification accuracy associated with each indiscernible DACV is obtained by computing the cardinality ratio of the corresponding β -lower approximate sets to the β -upper approximate sets. In the present example, the classification accuracies are therefore equal to ${}_{\beta}\alpha_1$ =2/2=1.000 and ${}_{\beta}\alpha_2$ = 2/2=1.000, respectively.

The FVM procedure then determines the multi-dimensional centroids of the β -lower approximate sets associated with each indiscernible DACV by calculating the mean attribute values (both conditional and decision) of all the instances within the corresponding sets. Thus, in the present example, the centroids of the lower approximate sets associated with the two indiscernible DACVs are obtained as $z_1' = mean(x \mid x \in {}_{\beta}\underline{R}(X), C_D(x) = 1) = mean(x \mid x \in \{x_1, x_2\}) = ((-1.40 - 1.60)/2, (1.30 + 1.20)/2, (1.05 + 0.95)/2, (-0.95 - 0.85)/2) = (-1.50, 1.25, 1.00, -0.90)$ and $z_2' = mean(x \mid x \in {}_{\beta}\underline{R}(X), C_D(x) = 2) = mean(x \mid x \in \{x_3, x_4\}) = (-2.00, 1.50, 0.55, -0.65)$, respectively.

Having determined the aggregative membership function values of all the instances, the classification accuracy, and the centroids of the β -lower approximate sets, the optimality of the discretization / classification outcome is evaluated using the MD index function (i.e., $MD(C) = \left(\frac{1}{C} \times \frac{E_1}{\overline{F}_{N_I}} \times \overline{D}_{N_I} \right)$). In describing the derivation of $\overline{F}_{N_I}$ (where $\overline{F}_{N_I} = \sum_{c=1}^{N_I} E_c'$), the following discussions arbitrarily consider the computation of $E_2^{'}$. (Note, that $E_1^{'}$ is computed in an identical manner.). The first instance in the dataset, x_1, has attribute values of $x_1(-1.40, 1.30, 1.05, -0.95)$. In addition, the centroid of the β -lower approximate sets associated with the second indiscernible DACV is given by $z_2'(-2.00, 1.50, 0.55, -0.65)$. As a result, $(x_1(a_1) - z_2'(a_1)) = (-1.40 - (-2.00)) = 0.60$, $(x_1(a_2) - z_2'(a_2)) = (1.30 - 1.50) = -0.20$,

$(x_1(d_1) - z_2'(d_1)) = (1.05 - 0.55) = 0.50$, and $(x_1(d_2) - z_2'(d_2)) = (-0.95 - (-0.65)) = -0.30$. Therefore, the vector of $x_{12} = x_1 - z_2'$ has the form $[x_{12}(a_1), x_{12}(a_2), x_{12}(d_1), x_{12}(d_2)] = [0.60, -0.20, 0.50, -0.30]$, and the corresponding norm is equal to $\|x_1 - z_2'\| = \sqrt{x_{12}(a_1)^2 + x_{12}(a_2)^2 + x_{12}(d_1)^2 + x_{12}(d_2)^2} = \sqrt{0.60^2 + (-0.20)^2 + 0.50^2 + (-0.30)^2} = 0.860$. Let the fuzzification parameter m' be specified as 2.0. Applying the notation $\|x_{j2}\| = {}_d\mu_{j2}^2(x_j(d)) \times \|x_j - z_2'\|$, the effect of instance x_1 on z_2', i.e., $\|x_{12}\|$, is obtained by multiplying $\|x_1 - z_2'\|$ by the square of the corresponding membership function, i.e., ${}_d\mu_{12}^2(x_1(D)) = 0.000^2 = 0.000$. Thus, $\|x_{12}\|$ has a value of 0.000. $\|x_{22}\|, \|x_{32}\|$ and $\|x_{42}\|$ are calculated using an identical procedure. The corresponding results are shown in Table 1(d). The value of E_2' is thus obtained as $E_2' = (\sum_{j=1}^{4} {}_d\mu_{j2}^2(x_j(d))\|x_j - z_2'\|) \Big/ {}_\beta\alpha_2 = (\sum_{j=1}^{4} \|x_{j2}\|) \Big/ {}_\beta\alpha_2 = (\|x_{12}\| + \|x_{22}\| + ... + \|x_{42}\|) / {}_\beta\alpha_2 =$ (0.000+0.000+0.088+0.095)/1.000=0.183. Utilizing an identical approach to that described above, the value of E_1' is obtained as 0.243. $\overline{F}_{N_I}$ is thus found to have a value of $\overline{F}_{N_I} = \sum_{c=1}^{2} E_c' = 0.426$.

the i-th instance	${}_d\overline{\mu}_{1i}(x_i(d))$	${}_d\overline{\mu}_{2i}(x_i(d))$
1	0.975	0.010
2	0.939	0.015
3	0.015	0.939
4	0.010	0.975

Table 1(c) Aggregative membership functions of instances (obtained using minimize peration)

x_j	z_c'	
j	c=1	c=2
1	0.126	0.000
2	0.117	0.000
3	0.000	0.088
4	0.000	0.095
$\sum_{j=1}^{4}\|x_{jc}\|$	0.243	0.183

Table 1(d)

Values of $\|x_{jc}\|$ $(= {}_d\overline{\mu}_{jc}^2(x_j(d)) \times \|x_j - z_c'\| / {}_\beta\alpha_c)$

Factor E_1 in the MD index function is a constant for a given dataset in which the instances belong to only one cluster. As a result, the attribute values of the centroid z_1 of the illustrative dataset can be obtained using the arithmetic mean function $mean(x \mid x \in \{x_i\}, i = 1,2,...,4)$ as ($((-1.40) + (-1.60) + (-1.95) + (-2.05))$, $(1.30 + 1.20 + 1.45 + 1.55)$ $(1.05 + 0.95 + 0.60 + 0.50)$, $((-0.95) + (-0.85) + (-0.70) + (-0.60))$)= $z_1(-1.750, 1.375, 0.775, -0.775)$. Based on the vector of centroid z_1 , it can be shown that $(x_1(a_1) - z_1(a_1)) = ((-1.40) - (-1.750)) = 0.350$, $(x_1(a_2) - z_1(a_2)) = (1.30 - 1.375) = -0.075$, $(x_1(d_1) - z_1(d_1)) = (1.05 - 0.775) = 0.275$, and $(x_1(d_2) - z_1(d_2)) = ((-0.95) - (-0.775)) = -0.175$. Therefore, the vector of $x_{11} = x_1 - z_1$ has the form $[x_{11}(a_1), x_{11}(a_2), x_{11}(d_1), x_{11}(d_2)] = [0.350, -0.075, 0.275, -0.175]$, and the corresponding norm is equal to $\|x_1 - z_1\| = \sqrt{x_{11}(a_1)^2 + x_{11}(a_2)^2 + x_{11}(d_1)^2 + x_{11}(d_2)^2} = \sqrt{0.350^2 + (-0.075)^2 + 0.275^2 + (-0.175)^2} = 0484$. Similarly, the norms

of $\|x_2 - z_1\|, \|x_3 - z_1\|$ and $\|x_4 - z_1\|$ are found to be 0.299, 0.286 and 0.476, respectively. The value of E_1 in the MD index function is then obtained by summing the norms of $\|x_j - z_1\|$ where $j = 1,2,...,4$, yielding a value of E_1 = 1.546.

The value of D'_{N_I} in the MD index function is acquired by calculating the maximum separation distance between the centroids of the lower approximate sets associated with the first and second indiscernible DACVs. In the present example, these centroids are given by z'_1(-1.50,1.25,1.00,–0.90) and z'_2(–2.00,1.50, 0.55,–0.65), respectively. Thus, the vector of $z_{12} = z'_1 - z'_2$ which maximizes the value of $D'_{N_I} = \max_{i,j=1}^{N_I} \|z'_i - z'_j\|$ has the form $[z_{12}(a_1), z_{12}(a_2), z_{12}(d_1), z_{12}(d_2)]$ = [0.50,-0.25,0.45,–0.25]. The corresponding norm is therefore equal to $\sqrt{0.50^2 + (-0.25)^2 + 0.45^2 + (-0.25)^2} = 0.76$.

Given the parameter values specified / derived above (i.e., $C = 2$, E_1 = 1.546, $\overline{F}_{N_I} = 0.426$ and $D'_{N_I} = 0.76$), the MD index function ($MD(C) = \left(\frac{1}{C} \times \frac{E_1}{\overline{F}_{N_I}} \times \overline{D}_{N_I} \right)$) returns a value of 1.379.

4 THE EVALUATION OF FVM-INDEX METHOD

In the stage of Data Processing / Statistical and Multivariate Analysis, we use a hybrid data-mining technique to construct an MCDM model to eliminate the uncertain instances from the original datasets.

4.1 A simple case study

This example was mainly used to evaluate the effect on the reliable decision-making rules obtained when the number of decision attributes were been increased.

code of instances	Condition Attributes			Decision Attributes	
	C1	C2	C3	D1	D2
1	1	1	1	1	1
2	1	1	2	1	1
3	1	2	1	2	2
4	1	2	1	2	2
5	1	2	2	2	1
6	2	2	1	2	1
7	2	2	1	2	2
8	2	2	1	2	2
9	2	2	2	2	2
10	2	2	2	1	2

Table 2 The approach of deleting instances in the synthetic dataset.

In Table 2, the example considers a hypothetical dataset in which each instance has three conditional attributes, i.e., a_1, a_2 and a_3, and two decision attributes, i.e., d_1 and d_2. The dataset is assumed to contain 10 instances, i.e., x_1 to x_{10}. In performing the clustering / classification process, an assumption is made that all of the attributes (both conditional and

decision) can be partitioned into two clusters. To explore the effect of the number of the decision attributes on the β-lower approximation, two single-attribute decision-making (FVS) methods and the FVM-index method was applied to this hypothetical dataset. When we are clustering the dataset, both the FVS- and FVM-index have to consider all conditional attributes, i.e., a_1, a_2 and a_3. However, the FVM-index method also has to consider two decision attributes, d_1 and d_2. But in the FVS-index methods, FVS_1 only has to consider the first decision attribute d_1; and FVS_2 has to consider d_2. The β-lower approximation sets were obtained using the FVM-, FVS_1- and FVS_2-index methods.

(1) The β-lower approximation sets obtained using the FVM-index method

The instances belong to β-lower approximation sets obtained using the FVM-index method were $\{x_1, x_2, x_3, x_4, x_5\}$. The sixth, the seventh, and the eighth instances were deleted because that the DACV of the sixth instance is {2, 1} while the DACV of the seventh and the eighth instances are {2,2} whereas the conditional attributes clusters vector in three of them are all {2,2,1}. Similarly, the ninth and the tenth instances were also deleted because that the DACV of the ninth instance is {2,2} while the DACV of the seventh and the tenth instance are {2, 1} whereas the conditional attributes clusters vector in two of them are all {2,2,2}. Therefore, three reliable-decision-making rules using the proposed FVM-index method are extracted from this dataset.

(2) The β-lower approximation sets obtained using the FVS_1 index method

The instances belong to β-lower approximation sets obtained using the FVS_1 index method were $\{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}$. Compared to the FVM-index method, the sixth, the seventh, and the eighth instances were kept in the β-lower approximation sets because that the DACV of all the three instance are {2} whereas the conditional attributes clusters vector in three of them are all {2,2,1}. However, the ninth and the tenth instances were still deleted because that the DACV of the ninth instance is {2} while the DACV of the seventh and the tenth instance are {1} whereas the conditional attributes clusters vector in two of them are all {2,2,2}.

(3) The β-lower approximation sets obtained using the FVS_2 index method

The instances belong to β-lower approximation sets obtained using the FVS_2 index method were $\{x_1, x_2, x_3, x_4, x_5, x_9, x_{10}\}$. Compared to the FVM-index method, the sixth, the seventh, and the eighth instances were still deleted because that the DACV of the sixth instance is {1} while the DACV of the seventh and the eighth instance are {2} whereas the conditional attributes clusters vector in two of them are all {2,2,1}. However, the ninth and the tenth instances were kept in the β-lower approximation sets because that the DACV of both the two instance are {2} whereas the conditional attributes clusters vector in two of them are all {2,2,2}.

From the above description, the instances belong to β-lower approximation sets obtained using the FVM-index method is lower than those of the FVS_1- and FVS_2- index methods (i.e., 5 .vs. 8 or 7). The inclusion of a greater number of decision attributes in the clustering process results in fewer reliable decision rules.

5 CONCLUSIONS

The proposed FVM-index method consists of Fuzzy Theory, VPRS Theory, and refined Huang index function. The method provides the means to determine the optimal number of attribute clusters within the dataset and the optimal classification accuracy. As for the consequences indicate, the proposed FVM-index method provides an effective means of

filtering inaccurate instances and extracting reliable-decision-making rules from datasets. The main conclusions of this current research are presented as follows:

The more attributes the instances contain, the more complicated inter-relationships between each attribute should be considered. Therefore, when the decision attributes increased, both the inter-relationships between attributes and the optimal clustering results will become much more complicated. In addition, when the decision attributes increase, the instances in the β-lower approximate sets will become lesser. In the future, the effectiveness of the proposed method will be confirmed by using Canonical Correlation Analysis.

Acknowledgements

This study was financially supported by the Research Grant NSC 102-2410-H-275 -006 - from Taiwan's Ministry of Science and Technology.

References

Berredoa, R. C., Ekelb, P.Y., Palharesc, R. M. (2005). Fuzzy preference relations in models of decision making, Nonlinear Analysis, 63, e735-e741.

Bezdek, J.C. (1981). Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York.

Bezdek, J.C. (1974). Cluster validity with fuzzy sets, J. Cybernet, 3, 58-74.

Chen, S. M., Lee, L.W. (2010). Fuzzy multiple attributes group decision-making based on the ranking values and the arithmetic operations of interval type-2 fuzzy sets, Expert Systems with Applications, 37, 824-833.

Dillon, William R., and Goldstein, M., Multivariate analysis. New York: Wiley, 1984.

Dunn, J.C. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. J. Cybernet, 3, 32-57.

Fan, Z.-P., Hu, G.-F., Xiao, S.i-H. (2004). A method for multiple attribute decision-making with the fuzzy preference relation on alternatives, Computers & Industrial Engineering, 46, 321-327.

Figueira, J.R., Liefooghe, A., Talbi, E.-G., Wierzbicki, A.P. (2010). A parallel multiple reference point approach for multi-objective optimization, European Journal of Operational Re-search, 205, 390-400.

Ghosh, A., Mishra, N.S., Ghosh, S. (2011). Fuzzy clustering algorithms for unsupervised change detection in remote sensing images, Information Sciences, 181, 699-715.

Guo, M., Yang J.-B., Chin, K.-S., Wang, H.-W. (2007). Evidential reasoning based preference programming for multiple attribute decision analysis under uncertainty, European Journal of Operational Research, 182, 1294-1312.

Hardoon, D. R., Szedmak, S., & Shawe-Taylor, J. (2004). Canonical correlation analysis: An overview with application to learning methods. Neural Computation, 16(12), 2639-2664.

Hotelling, H., Relations between two sets of variates. Biometrika 28, 321-377.

Huang, K. Y. (2009). Application of VPRS model with enhanced threshold parameter selection mechanism to automatic stock market fore-casting and portfolio selection. Expert Systems With Applications, 36(9), 11652-11661.

Huang, K.-Y. (2010). Applications of an enhanced cluster validity index method based on the Fuzzy C-means and Rough Set Theories to partition and classification. Expert Systems with Applications. 37(12), 8757-8769.

Ishibuchi, H., Murata, T. (1998). A multi-objective genetic local search algorithm and its application to flowshop scheduling. IEEE Transactions on Systems, Man and Cybernetics, 28, 392-403.

Johnson, R. A., & Wichern, D. W. , Applied multivariate statistical analysis. Upper Saddle River, NJ: Prentice hall, 2002.

Lioua, J. J. H. & Tzeng,G.-H. (2012). Comments on “Multiple criteria decision making (MCDM) methods in economics: an overview”, Technological and Economic Development of Economy, 18, 672-695.

Pakhira, M.K., Bandyopadhyay, S., Maulik, U. (2004). Validity index for crisp and fuzzy clusters, Pattern Recognition, 37, 487-501.

Roy, A., Pal, S.K. (2003). Fuzzy discretization of feature space for a rough set classifier, Pattern Recognition Letters, 24, 895–902.

Steuer, R.E. (1986). Multiple Criteria Optimization: Theory, Computation, and Application. John Wiley & Sons, Chichester, UK,.

Saaty, T.L. (1980). The Analytic Hierarchy Process, McGraw-Hill, New York.

Thompson, B., ed. Canonical correlation analysis: Uses and interpretation. No. 47. Sage, 1984.

Triantaphyllou, E. & Mann, S. H. (1989). An examination of the effectiveness of multi-dimensional decision-making methods: A decision-making paradox，Decision Support Systems, 5(3), 303-312.

Tseng, C.-C., Hong, C.-F., Chang, H.-L. (2008). Multiple Attributes Decision-Making Model for Medical Service Selection: An AHP Approach，Journal of Quality，15(2)，155-165.

Tsoukalas, L. H., Uhrig, R. E. (1997). Fuzzy and Neural Approaches in Engineering, John Wiley & Sons, Inc..

Tzeng, G.-H. & Huang, J.-J., Multiple Attribute Decision Making: Methods and Applications, Chapman and Hall/CRC, 2011.

Wang, W., Zhang, Y. (2007). On fuzzy cluster validity indices, Fuzzy Sets and Systems, 158, 2095-2117.

Wu, Q., Law, R. (2010). Complex system fault diagnosis based on a fuzzy robust wavelet support vector classifier and an adaptive Gaussian particle swarm optimization, Information Sciences. 180, 4514-4528.

Yager, R. R. (2006). An extension of the naive Bayesian classifier, Information Sciences, 176, 577-588.

Yeh, C.-H., Chang, Y.-H. (2009). Modeling subjective evaluation for fuzzy group multicriteria decision making, European Journal of Operational Research, 194, 464-473.

Zadeh, L.A. (1965). Fuzzy sets, Inf. Control, 8, 338-353.

Ziarko, W (1993). Variable precision rough set model. Journal of Computer and System Sciences, 46, 39-59.