A Hybrid Approach to Continuous Valued Datasets Classifying based on Particle Swarm Optimization, Variable Precision Rough Set Theory and Modified Huang-index Function

Kuang Yu Huang
Department of Information Management
Ling Tung University
#1 Ling Tung Road, Taichung City 408
Taiwan
kyhuang@mail.ltu.edu.tw

Abstract: - This paper proposed a new hybrid method, designated as PSOVPRS-index method, for partitioning and classifying continuous valued datasets based on particle swarm optimization (PSO) algorithm, Variable Precision Rough Set (VPRS) theory and a modified form of the Huang-index function. In contrast to the Huang-based index method which simply assigns a constant number of clusters to each attribute and in which the Rough Set (RS) theory is applied, this method could not only cluster the values of the individual attributes within the dataset and achieves both the optimal number of clusters and the optimal classification accuracy, but also extends the applicability of classification using VPRS theory. The validity of the proposed approach is investigated by comparing the classification results obtained for a real-world dataset containing stock market information with those obtained by PSORS-index method and pseudo-supervised decision-tree classification method. There is good evidence to show that the proposed PSOVPRS-index method not only has a better classification performance than the considered methods, but also achieves a more reliable basis for the extraction of decision-making rules.

Key-Words: - Particle Swarm Optimization, Variable Precision Rough Set theory, PSOVPRS-index method, classification, discretization, pseudo-supervised classification method

1 Introduction

Classification, discretization and dimensionality reduction are key challenges in the pattern recognition, machine learning and data mining fields. Classification is a task of assembling data with multiple attributes into relevant categories and provides an invaluable means of uncovering the implicit knowledge within a dataset. In the last few decades, several articles have been devoted to the study of the classification algorithms, including decision-tree algorithms such as ID3 [1], rule-based algorithms such as CN2 [2], Bayesian classifiers [3], conformal predictors [4], back-propagation networks [5], support vector machines [6, 7], and so forth. All of these schemes have their respective merits and have found widespread use in a diverse range of applications, including weather prediction, manufacturing process planning, medical diagnosis, and so on. However, such methods can not deal effectively with datasets having no categorical values specified or datasets characterized by uncertain or missing information. Thus, a best possible solution to the problem of interest is applied the Rough Set (RS) theory [8-11] to classify the continuous valued datasets.

RS theory was first introduced more than twenty years ago [8] and has emerged as a powerful technique for the automatic classification of records [9] in such fields as machine learning, forecasting, knowledge acquisition, decision analysis, knowledge discovery, and pattern recognition. However, the ability of RS techniques to correctly classify a dataset relies upon the availability of complete and certain information. To perform a classification operation with a controlled degree of uncertainty or misclassification error is beyond the ability of the RS approach [12]. To extend RS theory to such classification applications, Variable Precision Rough Set (VPRS) proposed by Ziarko is a methodology in which the records within the dataset were analyzed and classified in terms of their statistical tendencies rather than their functional patterns [12, 13]. In VPRS theory, the uncertain nature of the information within the dataset of interest is handled using the concept of β-lower and β-upper approximate sets. However, the performance of VPRS models [14] is basically

resulting from the quality of the original clustering results. Attributes clustering must be performed in prior to conduct a continuous valued dataset classification, and correct partitioning is the prelude to available classifications. In general, the problem of evaluating the optimality of the clustering results obtained for a particular dataset is referred to as the cluster validity problem [15]. Many methods have been proposed for assessing the validity of the clustering results obtained using fuzzy clustering schemes [16], such as classification entropy [17] and the using matrix U methods [15, 18, 19]. Accordingly, when classifying continuous valued datasets with uncertain or missing information, it is preferable to utilize Variable Precision Rough Set (VPRS) theory for classification purposes, and to integrate the VPRS model with some form of cluster generation / cluster index evaluation procedure such that the optimal discretizing solution can be obtained. In a recent research, Huang [20] proposed a Huang-based index method which simply assigns a constant number of clusters to each attribute. One major question which exists in Huang-index clustering method is in adopting an attribute-based clustering approach, how one can determine the optimal number of clusters for each conditional and decision attribute values of the records. In order to optimize the number of clusters per conditional and decision attribute, it is necessary to integrate the clustering mechanism with some form of optimization technique.

Particle Swarm Optimization (PSO), inspired by the natural phenomena of bird flocking and fish schooling, provides a powerful technique for solving a wide range of classification and optimization problems [21-24]. Accordingly, PSO is an ideal tool for solving the problem considered in this study, namely that of determining the optimal number of clusters per conditional and decision attribute for a continuous valued dataset. In the proposed approach, designated as the PSOVPRS-index method, the PSO algorithm is integrated with a modified form of the Huang-index method [20] referred to as the FV-index method. The FV-index method comprises a fuzzy clustering scheme, a VPRS classification model in which the optimal threshold parameter β is determined using the method presented by Huang [11], and a so-called VM-index function, which evaluates the optimality of the discretization / classification results in terms of both the number of clusters within the dataset and the accuracy of β-approximation. Broadly speaking, the PSOVPRS-index method provides the means to solve the following problems for continuous valued datasets: (1) discretizing the continuous values of each attribute within the dataset; (2) determining both the optimality of the discretization results and the optimal number of clusters per attribute; (3) and extending the applicability of classification using VPRS theory.

The remainder of this paper is organized as follows. In the next section, we present the fundamental principles of the VPRS theory, VM-index function and PSO, respectively. In Section 3, we interpret the combination of the concepts to form the proposed PSOVPRS-index method. In Section 4, we compare the performance of the proposed method with those of the PSORS-index method [25] and pseudo-supervised decision-tree classification method. The paper concludes in Section 5 with some brief remarks and indicates the intended direction of future research.

2 Review of Related Methodologies

2.1 Index function $I_{\max}$ [20]

Assume that each record x_i in the dataset has 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 record x_i belongs. Here $C_{a_l}(x_i)$ is given by

$$C_l(x_i) = I_{\max}(\mu_j(x_i(a_l))) = \text{Index}(\max(\mu_j(x_i(a_l)))) \quad for \quad 1 \le l \le m, 1 \le i \le n \quad (1)$$

where $\mu_j(x_i(a_l))$ is the membership function values of the l-th attribute of x_i; $\max(\mu_j(x_i(a_l)))$ returns the maximum of these membership functions values; $\text{Index}(\max(...))$ will return the index of the cluster associated with the output of $\max(...)$; Therefore, $I_{\max}(\mu_j(x_i(a_l)))$ 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 Fundamental principles of VPRS theory

The VPRS operates on what may be described as a knowledge-representation system, or information system [12]. The basic principles and notations of information systems (S) and the application of VPRS theory to the processing of such systems are described in the sections below.

2.2.1 β-lower and β-upper approximate sets

For a given dataset, any records which are indistinguishable from one another when evaluated using a particular subset of all the attributes define an equivalence or indiscernibility relationship. In VPRS theory, this indiscernibility concept is handled using 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 and $X \subseteq U$ and $R \subseteq A$. Generally speaking, the attributes in set A can be partitioned into a set of conditional attributes $C \neq \phi$ and a set of decision attributes $D \neq \phi$, i.e. $A = C \cup D$ and $C \cap D = \phi$. For each attribute, $q \in A$, V_q represents the domain of q, i.e. $V = \cup V_q$. 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 proposed VPRS method utilizes the systematic method presented by the current author in [11] to determine an appropriate value of the threshold parameter β, i.e., the value of β at which a certain proportion of the records in a particular conditional class are classified into the same decision class. When processing an information system using a VPRS model with $0.5 < \beta \le 1$, the objective is to identify the β-lower and β-upper approximate sets associated with each cluster of the decision attribute. In general, the β-lower approximation of sets $X \subseteq U$ and $P \subseteq C$ can be expressed as

$$\begin{aligned} {}_{\beta}\underline{R}_P(X) &= \{x \in U : P(X/[x]_P) \ge \beta\} \\ &= \cup\{[x]_P : P(X/[x]_P) \ge \beta\} \end{aligned} \tag{2}$$

Similarly, the β-upper approximation of sets $X \subseteq U$ and $P \subseteq C$ is given by

$$\begin{aligned} {}_{\beta}\overline{R}_P(X) &= \{x \in U : P(X/[x]_P) > 1-\beta\} \\ &= \cup\{[x]_P : P(X/[x]_P) > 1-\beta\} \end{aligned} \tag{3}$$

Note that $P(X/Y) = |X \cap Y|/|Y|$ if $|Y| > 0$, and $P(X/Y) = 1$ otherwise. Note also that $|X|$ denotes the cardinality of set X. In the particular case of $\beta = 1$, ${}_{\beta}\underline{R}_P(X)$ and ${}_{\beta}\overline{R}_P(X)$ are equivalent to the lower and upper approximate sets in RS theory. In other words, the VPRS model reverts to the traditional RS model.

2.2.2 Accuracy of VPRS classification results

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

$${}_{\beta}\alpha_c = \left|{}_{\beta}\underline{R}_P(X)\right| \Big/ \left|{}_{\beta}\overline{R}_P(X)\right| \tag{4}$$

where $X = \{x : C_d(x) = c, \forall x \in U\}$, and $|{}_{\beta}\underline{R}_P(X)|$ and $\left|{}_{\beta}\overline{R}_P(X)\right|$ are the cardinalities of the β-lower and β-upper approximate sets, respectively, when classifying the elements (x) in terms of the c th cluster of the decision attribute d.

2.3 Overview of Huang and VM cluster optimization index functions

2.3.1 Accuracy of VPRS classification results

The Huang-based index method in which the RS is applied assigns a constant number of clusters to each attribute and is applied to optimize both the number of clusters within the dataset and the corresponding classification accuracy. This function has the form:

$$H(N_d, \alpha_C) = \left(\frac{1}{\mathrm{N}_d} \times \frac{\overline{E_1}}{F'_{N_d}} \times D'_{N_d} \right) \tag{5}$$

where N_d is the number of clusters assigned to the decision attributes and α_c is the corresponding classification accuracy when evaluated in terms of the c th cluster of the decision attribute d. $\overline{E_1}$ is a constant for a given dataset and is set in such a way as to prevent the second term from vanishing. In addition, $F'_{N_d} = \sum_{c=1}^{N_d} E'_c$, E'_c is obtained by accumulating the value of E'_c for each cluster of the decision attribute (d), where E'_c is given by

$$E'_c = \sum_{j=1}^{n} \overline{\mu}_{cj}^{\,m'}(x_j(d)) \left\| x_j - z'_c \right\| \Big/ \alpha_c \tag{6}$$

in which $\overline{\mu}_{cj}(x_j(d))$ is the membership function of record x_j in the c th cluster of the decision attribute d, and z'_c is the multi-dimensional centroid of the lower approximate sets in terms of the c th cluster of the decision attribute d and is obtained by computing the mean values of the conditional and decision attribute values of each record within the corresponding sets. Furthermore, m' is the fuzzification parameter and n is the total number of records in the dataset. Finally, the value of D'_{N_d} is equal to the maximum separation distance amongst the centroids of all the lower approximate sets in terms of the different clusters of the decision attribute,

$$\text{i.e., } D'_{N_d} = \max_{i,j=1}^{N_d} \left\| z'_i - z'_j \right\| \tag{7}$$

where z'_i , z'_j are the multi-dimensional centroids of the lower approximate sets in terms of the ith and jth clusters, respectively, of the decision attribute d .

2.3.2 VM-index function

In contrast to the Huang-index function which is based on the RS classification approach, the VM-index function applies the VPRS classification scheme to extend applicability of the Huang-index function to handle the uncertain information system. The VM-index function proposed in this study has the form:

$$VM(C_M, {}_\beta\alpha_c) = \left(\frac{1}{C_M} \times \frac{\overline{E_1}}{{}_\beta F'_{N_d}} \times D'_{N_d}\right) \quad (8)$$

where C_M is the arithmetic mean of numbers of clusters of individual attributes; N_d is the number of clusters of the conditional and decision attributes, and $\overline{E_1}$ is constant and the same as defined in Huang-based index method; ${}_\beta F'_{N_d}$ ($= \sum_{c=1}^{N_d} {}_\beta E'_c$) is obtained by aggregating the value of ${}_\beta E'_c$ for each cluster of the decision attribute (d), where ${}_\beta E'_c$ is given by

$${}_\beta E'_c = \sum_{j=1}^{n} \overline{\mu}_{cj}^{m'}(x_j(d)) \left\| x_j - z'_c \right\| \Big/ {}_\beta\alpha_c \quad (9)$$

in which ${}_\beta\alpha_c$ is the accuracy of VPRS classification when evaluated in terms of the c-th cluster of the decision attribute, $\mu_{cj}(x_j(d))$ is the membership function of record x_j in the c-th cluster of the decision attribute d and z'_c is the multi-dimensional centroid of the lower approximate sets associated with the c -th cluster of the decision attribute d and is obtained by computing the mean values of the conditional and decision attributes of each record within the corresponding sets, m' is the fuzzification parameter, and n is the total number of records in the dataset; finally, the value of D'_{N_d} is equal to the maximum separation distance amongst the centroids of all the lower approximate sets associated with the different clusters of the decision attribute, i.e. $D'_{N_d} = \max_{i,j=1}^{N_d} \left\| z'_i - z'_j \right\|$. Note that the value of D'_{N_d} is upper bounded by the maximum separation distance amongst all possible pairs of records in the dataset.

Note that parameter ${}_\beta F'_{N_d}$ in the VM-index function differs slightly from parameter F'_{N_d} in the Huang-index function described in Section 2.3.1. The value of ${}_\beta F'_{N_d}$ depends on ${}_\beta\alpha_c$ in VM-index function, while the value of F'_{N_d} depends on α_c in Huang-index function.

2.4 Particle swarm optimization (PSO) theory

In PSO [25], the positions of particles within the swarm describe candidate solutions for the n -dimensional problem under consideration, and the movements of the particles describe the search for a better solutions [22, 23]. Let the position of the k th particle be described as

$$x_{k_j} = (x_{k_1}, x_{k_2}, \ldots x_{k_n}) \quad (10)$$

and let its velocity be represented by

$$v_{k_j} = (v_{k_1}, v_{k_2}, \ldots v_{k_n}) \quad (11)$$

, $j = 1,2,...,n$. During the search process, the particles successively adjust their positions in accordance with two features, namely their personal best position and the global best position. The personal best position

$$P_k = (P_{k_1}, P_{k_2}, \ldots P_{k_n}) \quad (12)$$

of particle k is defined as the position of this particle which yields the highest fitness value to date. Meanwhile, the global best position

$$P_g = (P_{g_1}, P_{g_2}, \ldots P_{g_n}) \quad (13)$$

is defined as the position which yields the greatest fitness value amongst all the particles' positions to date.

In this study, the velocity and position of the k th particle is updated using the "constrict factor method" proposed by Clerc [26], i.e.,

$$v_{k_j}(t+1) = \varphi \left(v_{k_j}(t) + C_1 r_1 \left(P_{k_j} - x_{k_j}(t) \right) + C_2 r_2 \left(P_{g_j} - x_{k_j}(t) \right) \right) \quad (14)$$

$$x_{k_j}(t+1) = x_{k_j}(t) + v_{k_j}(t+1) \quad (15)$$

, $j = 1,2,...,n$.

where r_1 and r_2 are elements from two uniform random sequences in the range $(0,1)$; $C_1(2.05)$ is the individual factor; C_2 (2.05) is the societal factor; and

$$\varphi = 2 \Big/ \left| 2 - C - \sqrt{C^2 - 4C} \right| \quad (16)$$

, where $C = C_1 + C_2$.

The positions of particles in each dimension are clamped to a maximum position $x_{\max}$. If the sum of velocities would cause the position of that dimension to exceed $x_{\max}$, which is a parameter specified by the user, then the position of that dimension is limited to $x_{\max}$.

In the attribute values clustering problem considered in this study, the particles represent candidate solutions for the number of clusters assigned to the corresponding attribute of datasets and have the form of a string of real-valued numbers, where the length of the string corresponds to the total number of attributes in the dataset, while each element , N_{a_i} , of the string represents the number of clusters assigned to the corresponding attribute by a random function. In other words, the particle has the form (N_{a_1}, N_{a_2} ..... N_{a_k}, N_d). In this form, k denotes the number of conditional attributes, N_{a_i} indicates the number of clusters assigned to the i th conditional attribute a_i , and N_d represents the number of clusters assigned to the decision attribute d .

3 PSOVPRS-index Method

The PSOVPRS-index method proposed in this study is used to extend applicability of the PSORS-index method [25], discretizes the values of the individual attributes within the dataset and achieves both the optimal number of clusters and the optimal classification accuracy. This method consists of a PSO and a FV-index method. In the FV-index method, the conditional and decision attribute values of the records in the dataset are fuzzified and discretized using the Fuzzy C-means (FCM) method in accordance with the cluster vectors given by the PSO and a rounding function specifying the number of clusters per attribute. Then, RS theory is first applied to determine the centroids of the lower approximate sets associated with each cluster of the decision attribute are determined by computing the mean conditional and decision attribute values of all the records within the corresponding sets. Secondly, VPRS theory is applied to determine the β-lower and β-upper approximate sets associated with each cluster of the decision attribute. Finally, the accuracy of VPRS classification of each cluster of the decision attribute is then computed as the cardinality ratio of the β-lower approximate sets to the β-upper approximate sets. The cluster centroids and accuracy of VPRS classification are then processed by a modified form of the Huang-index function, designated as the VM-index function, in order to determine the optimality of the discretization/classification results. In the event that the termination criteria are not satisfied, the PSO modifies the initial population of cluster vectors and the FV-index, comprising FCM, VPRS and VM-index function, procedures are repeated. The entire process is repeated iteratively until the termination criteria are satisfied. The maximum value of the VM cluster validity index is then identified, and the corresponding cluster vector is taken as the optimal classification result.

In accordance with the PSOVPRS-index method, each attribute of element (X_i) in U is mapped to an appropriate cluster amongst all of the clusters associated with the corresponding conditional attribute ($C_1 \sim C_n$) or decision attribute (d). The detailed parameters of the PSOVPRS-index function are presented in the following section. Table 1 summarizes the major components of the PSOVPRS-index function and the Huang-index function in order to emphasize the differences between them.

3.1 Details of proposed PSOVPRS-index method

Figure 1 illustrates the basic framework of the proposed PSOVPRS-index method. The details of each processing step are described in the following paragraphs.

Step 1: Generate PSO particles

As described in Section 2.4, the particles in the PSO algorithm have the form (N_{a_1}, N_{a_2} ..... N_{a_k}, N_d), where N_{a_i} indicates the number of clusters assigned to the i th conditional attribute a_i , and N_d represents the number of clusters assigned to the decision attribute d . The values of N_{a_i} and N_d are assigned by the random function. The number of decision attributes is indicated by default as one and the number, k , of conditional attributes is indicated in advance by the user. The PSO algorithm initializes by generating an initial population of $P = 40$ random candidate solutions and setting the specified number of iterations $M = 100$, where the values of each element of the particles is limited to the interval [2, $N_{\max}$]. This interval bounds the search space of the solution procedure for each attribute. That is, the minimum permissible value of the rounding function for each attribute is specified

as 2, while the maximum permissible values of the rounding function is specified as $N_{\max}$. Note that the upper bound value of $N_{\max}$ was specified by applying the Huang-based index method to a dataset with 2 conditional attributes and 1 decision attribute, and was used to provide a satisfactory classification performance. In order to ensure that the values of each element within the particle has a positive integer value derived from a rounding function, i.e.

$$\underline{N_{a_i}} = floor\left(N_{a_i}\right) \quad (17)$$

is applied as random values of N_{a_i} and N_d had determined for each candidate solution within the specified search range. So, the each number of clusters assigned to the i th conditional attribute a_i, and decision attribute d round down into the form ($\underline{N_{a_1}}$, $\underline{N_{a_2}}$ ..... $\underline{N_{a_k}}$, $\underline{N_d}$).

Step 2: Fuzzify attributes of dataset using FCM method

In this step, a continuous valued dataset can be converted into an equivalent fuzzy dataset using the Fuzzy C-Means clustering method.

Step 3: Assign each attribute of records to appropriate conditional or decision attribute clusters

Using the index function given in Section 2.1, each conditional or decision attribute cluster to which each attribute of each record belong is determined.

Step 4: Identify VPRS approximate sets and compute corresponding accuracy of VPRS classification

Having mapped the attribute values of all the records to the appropriate conditional or decision attribute clusters, the β -lower and β -upper approximate sets associated with each cluster c of the decision attribute d are extracted. The accuracy of VPRS classification associated with each cluster of the decision attribute is then obtained by calculating the cardinality ratio of the corresponding β -lower approximate sets to the β -upper approximate sets.

Step 5: Compute centroids of lower approximate sets associated with each cluster of the decision attribute

Using RS theory, the multi-dimensional centroids of the lower approximate sets associated with each cluster of the decision attribute d are obtained by computing the mean attribute values (both conditional and decision) of all of the records within the corresponding lower approximate sets.

Step 6: Determine value of VM-index function

Having determined the number of clusters per attribute, the membership function values of all the attributes of all the records, the accuracy of VPRS classification and the centroids of the lower approximate sets, the optimality of the discretization and classification solution is analyzed using the VM-index function.

Step 7: Compute fitness value

In the existent case, each particle specifies a possible number of clusters for each conditional and decision attribute, and the aim of the PSO optimization approach is to make sure the number of attribute clusters which optimizes both the separation of the clusters in the dataset and the corresponding accuracy of VPRS classification. Hence, in analyzing the relative quality of each potential clustering / classifying solution, the fitness of the solution is defined as the negative value of the corresponding VM-index function. In other words, the objective of the PSO approach is to specify the number of clusters for each conditional and decision attribute which provides the minimum fitness value (i.e. the maximum value of the VM-index function).

Step 8: Examine whether or not the termination criteria are satisfied

Having calculated the values of the VM-index function for each of the 40 particles in the current position, a check is made to see whether or not the termination criteria are satisfied (e.g. “is the fitness values of all particles are the same?”, “has the simulation run time reached the specified value?”, “have the specified number of iterations been evolved?”, and so on). If the termination criteria are not satisfied, the PSO creates a new candidate solutions using the updating operations of velocity and position described in Section 2.4. The FV-index computation approaches described in the steps

above are then repeated in order to identify the optimal solution in the new population. Once the termination criteria are satisfied, the iteration approach ends.

Step 9: Recognize value of VM cluster validity index

Once the termination criteria have been satisfied, the particle in the current time which yields the maximum value of the VM-index function is recognized. The corresponding value of the VM-index function is then recognized as the VM cluster validity index for the clustering / classification problem.

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

This section illustrates the derivation of the VM-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 one decision attributes, d. Let the four instances be defined as $x_1(0.90,0.30,-0.75)$, $x_2(1.10,0.20,-0.65)$, $x_3(1.45,0.45,-0.30)$ and $x_4(1.55,0.55,-0.20)$, respectively. In accordance with the PSOVPRS-index method, a PSO algorithm is first applied to initialize a set of random candidate solutions which indicate the numbers of clusters assigned to the conditional and decision attributes. Suppose that each conditional and decision attribute is partitioned into 2 clusters. Then, the continuous data in the hypothetical dataset are discretized using the FCM technique. The membership function values of each attribute of each instance are summarized in Table 2(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 2(b). As shown, the discretized vectors of the four instances x_i (I_{a_1}, I_{a_2}, I_d) have the form $x_1(2,2,2)$, $x_2(2,2,2)$, $x_3(1,1,1)$, and $x_4(1,1,1)$, respectively.

The upper and lower approximate sets associated with each cluster of the decision attribute are calculated in accordance with the formulation given in Section 2.3.2 of Ref [20] and are also shown in Table 2(b). Moreover, the threshold parameter β associated with first and second clusters of the decision attribute are determined in accordance with the procedure given in Section 2.2.2 of Ref [11] and are 0.974 and 0.939, respectively. Thus, the β-upper and β-lower approximate sets obtained using VPRS are the same as the upper and lower approximate sets obtained using RS. The accuracy of VPRS classification associated with each cluster of the decision attribute 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 α_1 =2/2=1.000 and α_2 = 2/2=1.000, respectively.

The PSOVPRS procedure then determines the multi-dimensional centroids of the lower approximate sets associated with each cluster of the decision attribute by calculating the mean attribute values (both conditional and decision) of all the instances within the corresponding sets using RS theory. Thus, in the present example, the centroids of the lower approximate sets associated with the two cluster of the decision attribute are obtained as $z_2' = mean(x \mid x \in \underline{R}(X), C_d(x) = 2) = mean(x \mid x \in \{x_1, x_2\}) = ((0.90+1.10)/2, (0.30+0.20)/2, (-0.75-0.65)/2) = (1.00, 0.25, -0.70)$ and $z_1' = mean(x \mid x \in \underline{R}(X), C_d(x) = 1) = mean(x \mid x \in \{x_3, x_4\}) = (1.50, 0.50, -0.25)$, respectively.

Having determined the membership function values of all the instances, the centroids of the lower approximate sets, and the accuracy of VPRS classification, the optimality of the discretization / classification outcome is evaluated using the VM-index function (i.e., $VM(C_{M,\beta}\alpha_c) = (\frac{1}{C_M} \times \frac{\overline{E_1}}{{}_\beta F'_{N_d}} \times D'_{N_d})$). In describing the derivation of ${}_\beta F'_{N_d}$ (where ${}_\beta F'_{N_d} = \sum_{c=1}^{N_d} {}_\beta E'_c$), the following discussions arbitrarily consider the computation of ${}_\beta E_1^{'}$. (Note, that ${}_\beta E_2^{'}$ is computed in an identical manner.). The first instance in the dataset, x_1, has attribute values of $x_1(0.90,0.30,-0.75)$. In addition, the centroid of the lower approximate sets associated with the first cluster of the decision attribute is given by $z_1'(1.50,0.50,-0.25)$. As a result, $(x_1(a_1) - z_1'(a_1)) = (0.90-1.50) = -0.60$, $(x_1(a_2) - z_1'(a_2)) = (0.30-0.50) = -0.20$, and $(x_1(d) - z_1'(d)) = (-0.75-(-0.25)) = -0.50$. Therefore, the vector of $x_{11} = x_1 - z_1'$ has the form

$[x_{11}(a_1), x_{11}(a_2), x_{11}(d)] = [-0.60, -0.20, -0.50]$, and the corresponding norm is equal to $\|x_1 - z_1'\| = \sqrt{x_{11}(a_1)^2 + x_{11}(a_2)^2 + x_{11}(d)^2} = \sqrt{(-0.60)^2 + (-0.20)^2 + (-0.50)^2} = 0.806$. Let the fuzzification parameter m' be specified as 2.0. Applying the notation $\|x'_{j1}\| = \mu_{1j}^2(x_j(d)) \times \|x_j - z_1'\|$, the effect of instance x_1 on z_1', i.e., $\|x'_{11}\|$, is obtained by multiplying $\|x_1 - z_1'\|$ by the square of the corresponding membership function value, i.e., $\mu_{11}^2(x_1(d)) = 0.010^2 = 0.000$. Thus, $\|x'_{11}\|$ has a value of 0.000. $\|x'_{21}\|$, $\|x'_{31}\|$ and $\|x'_{41}\|$ are calculated using an identical procedure. The corresponding results are shown in Table 2(c). The value of ${}_\beta E_1'$ is thus obtained as ${}_\beta E_1' = (\sum_{j=1}^{4} \mu_{1j}^2(x_j(d)) \|x_j - z_1'\|) \Big/ {}_\beta \alpha_1 = (\sum_{j=1}^{4} \|x'_{j1}\|) \Big/ {}_\beta \alpha_1 = (\|x'_{11}\| + \|x'_{21}\| + ... + \|x'_{41}\|) / {}_\beta \alpha_1 = (0.000+0.000+0.084+0.085)/1.000=0.169$. Utilizing an identical approach to that described above, the value of ${}_\beta E_2'$ is obtained as 0.239. ${}_\beta F'_{N_d}$ is thus found to have a value of ${}_\beta F_2' = \sum_{c=1}^{2} {}_\beta E_c' = 0.408$.

Factor $\overline{E_1}$ in the VM-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 $((0.90 + 1.10 + 1.45 + 1.55), (0.30 + 0.20 + 0.45 + 0.55), ((-0.75) + (-0.65) + (-0.30) + (-0.20))) = z_1(1.25, 0.375, -0.475)$.

Based on the vector of centroid z_1, it can be shown that $(x_1(a_1) - z_1(a_1)) = (0.90 - 1.250) = -0.350$, $(x_1(a_2) - z_1(a_2)) = (0.30 - 0.375) = -0.075$, and $(x_1(d) - z_1(d)) = (-0.75 - (-0.475)) = -0.275$. Therefore, the vector of $x_{11} = x_1 - z_1$ has the form $[x_{11}(a_1), x_{11}(a_2), x_{11}(d)] = [-0.350, -0.075, -0.275]$, and the corresponding norm is equal to $\|x_1 - z_1\| = \sqrt{x_{11}(a_1)^2 + x_{11}(a_2)^2 + x_{11}(d)^2} = \sqrt{(-0.350)^2 + (-0.075)^2 + (-0.275)^2} = 0451$. Similarly, the norms of $\|x_2 - z_1\|$, $\|x_3 - z_1\|$ and $\|x_4 - z_1\|$ are found to be 0.289, 0.276 and 0.443, respectively. The value of $\overline{E_1}$ in the VM-index function is then obtained by summing the norms of $\|x_j - z_1\|$ where $j = 1,2,...,4$, yielding a value of $\overline{E_1} = 1.460$.

The value of D'_{N_d} in the VM-index function is obtained by calculating the maximum separation distance between the centroids of the lower approximate sets associated with the first and second clusters of the decision attribute. In the present example, these centroids are given by $z_1'(1.50, 0.50, -0.25)$ and $z_2'(1.00, 0.25, -0.70)$, respectively. Thus, the vector of $z_{12} = z_1' - z_2'$ which maximizes the value of $D'_{N_d} = \max_{i,j=1}^{N_d} \|z_i' - z_j'\|$ has the form $[z_{12}(a_1), z_{12}(a_2), z_{12}(d)] = [0.50, 0.25, 0.45]$. The corresponding norm is therefore equal to $\sqrt{0.50^2 + 0.25^2 + 0.45^2} = 0.718$.

Given the parameter values specified / derived above (i.e., $C_M = 2$, $\overline{E_1} = 1.460$, ${}_\beta F_2' = 0.408$ and $D'_{N_d} = 0.718$), the VM-index function $(VM(C_M, {}_\beta \alpha_c) = (\frac{1}{C_M} \times \frac{\overline{E_1}}{{}_\beta F'_{N_d}} \times D'_{N_d}))$ returns a value of 1.284.

4 Performance evaluation of PSOVPRS-index Method

The validity and effectiveness of the proposed PSOVPRS-index method is evaluated by an illustrative example relating to electronic stock data extracted from the financial database maintained by the Taiwan Economic Journal (TEJ) [9, 25] for the first quarter of 2006. This database comprises 53 financial indices (attributes) for each stock item (instance). However, for simplicity, the performance evaluations conducted in this present case were restricted to just 6 conditional attributes (i.e., (i) Business Profit Rate, (ii) Pretax Income %, (iii) Net Nonop.Inc./Rev, (iv) PS-Pre_Tax Income, (v) Oper.Income/Capital, and (vi) Pre Tax Income/Capital) and 1 decision attribute (i.e., EPSNet Income). A total of 307 records were obtained (See Table 3 for indicative values of each index for a selected subset of these 307 records) as the records for which some of the data was incomplete had deleted.

In performing the evaluations, the effectiveness of the proposed method is explored by comparing the classification results with the results obtained from PSORS-index method and pseudo-supervised classification method. The PSOVPRS-index method

provides the means to discretize the continuous values of the individual attributes within a dataset and to classify datasets in which the records do not provide any class information. In contrast, supervised classification methods cluster attributes based on a consideration of class information. There are currently no classifiers available for the supervised classification of datasets with no class information. Therefore, it is impossible to establish a direct comparison between the classification performance obtained by the PSOVPRS-index method and those obtained from a supervised method. Accordingly, in this illustrative example, the classification performance of the PSOVPRS-index method is compared with those of pseudo-supervised decision-tree classification method, in which pseudo-class information is added to a dataset which initially lacks class information. The pseudo-class information is obtained by applying the PSOVPRS-index method to the target dataset in order to identify the optimal number of clusters for the decision attribute. The $I_{\max}$ function presented in Section 2.1 is then used to acquire the appropriate decision attribute cluster for each record in the dataset. The resulting cluster index is then treated as pseudo-class information for the record. Meanwhile, the classification performance of the PSOVPRS-index method is also compared with those of PSORS-index method using in [25] in which VPRS classification module was replaced by the conventional RS classification model. In this illustrative example, the PSOVPRS-index method, PSORS-index method, and the pseudo-supervised decision-tree classification method are used to classify training and testing datasets based upon a common 10-fold subsample of the stock market dataset. The optimal number of clusters for the decision attribute in this dataset is equal to 15, and thus the pseudo-class information added to the dataset to facilitate discretizing using the decision-tree classification method has a value in the interval [1, 15]. A common k-fold subsample (k=10) was used to confirm the performance of a classification method. Of the k subsamples, one subsample was retained for use as validation data in testing the method, while the remaining $k-1$ subsamples were used as training data.

The classification performance of the three methods is evaluated in terms of the classification accuracy (CA). For the case of the PSOVPRS- (or PSORS-) index method, CA is defined as the ratio of the total cardinality of the β-lower (or lower) approximation sets associated with each cluster of the decision attribute to the total number of samples in the dataset, i.e. $\sum_{c=1}^{N_d} \left| {}_{\beta}\underline{R}_P(X) \right| \Big/ |U|$ (or $\sum_{c=1}^{N_d} \left| \underline{R}_P(X) \right| \Big/ |U|$). For the pseudo-supervised decision-tree classification method, the CA is defined as the ratio of the number of records for which the measured class information is identical to the added pseudo-class information to the total number of records in the dataset. First, to compare the CA obtained for each training data and testing data through the PSOVPRS-index method with those through PSORS-index method [25], the CA obtained for each training data and testing data through the PSOVPRS-index method is higher than those through the PSORS-index method and pseudo-supervised decision-tree classification method, respectively, as shown in Table 4. Meantime, the average CA and the deviation of the CA obtained for the training and testing datasets by the PSOVPRS-index method, PSORS-index method and pseudo-supervised decision-tree classification method are shown in Table 4. It can be found that the PSOVPRS-index method produces an average CA of 0.79 for the training dataset and 0.99 for the testing dataset. In contrast, the PSORS-index method produces an average CA of 0.74 for the training dataset and 0.97 for the testing dataset, and the pseudo-supervised decision-tree classification method produces average CAs of 0.29 and 0.13 for the training dataset and testing dataset, respectively. In other words, the average CA obtained by the PSOVPRS-index method is higher than those obtained by the PSORS-index method and pseudo-supervised decision-tree classification method for both datasets, respectively. In addition, it is seen that the lowest CA values obtained by the PSOVPRS-index method for the training and testing datasets (i.e. 0.74 and 0.87, respectively) are higher than or equal to those obtained by the PSORS-index method (i.e. 0.68 and 0.87, respectively). Meanwhile, the lowest CA results obtained by the PSOVPRS-index method for the training and testing datasets are also higher than those obtained by the pseudo-supervised decision-tree classification method. Thus, the performance of the PSOVPRS-index method in optimizing the classification accuracy using a VPRS classification model is superior to those of the PSORS-index method in which the RS classification method is applied and pseudo-supervised classification decision-tree method in which a pseudo number of clusters is assigned to the decision attribute, respectively.

5 Conclusion

This study has presented a method designated as the PSOVPRS-index method for clustering and classifying complex, real-world datasets. This method is based on a PSO, VPRS theory and a VM-index function. The method provides the means to determine the optimal number of attribute clusters within the dataset and the optimal accuracy of VPRS classification. It should be conclude, from what has been said above, that:

(1) The PSOVPRS-index method is applicable to continuous value datasets in which the records do not provide any class information and may be imprecise and uncertain. Therefore, it is impossible to establish a direct comparison between the classification results of the PSOVPRS-index method and those of supervised methods since supervised methods depend on categorical information to cluster the attributes. However, it has been shown that the accuracy of VPRS classification of the PSOVPRS-index method is better than those of pseudo-supervised decision-tree classification method when applied to a dataset to which pseudo-class information is added to each record in order to facilitate classification.

(2) Applying a cross validation method to examine the accuracy of VPRS classification of the PSOVPRS-index method, the VRPS β-lower approximate set contains a greater number of instances than the RS lower approximate set. Therefore, the classification accuracy (CA) obtained for each training data and testing data through the PSOVPRS-index method is higher than those obtained through the PSORS-index method. In other words, the PSOVPRS-index method provides an extended applicability of classification using VPRS theory.

Overall, the evaluation results given in this research have confirmed that the proposed PSOVPRS-index method provides a practical tool for optimizing both the number of clusters of attributes and the accuracy of VPRS classification when applied to the clustering/classification of complex, real-world datasets. Consequently, the proposed PSOVPRS-index method will be used as the basis for an automatic portfolio selection mechanism designed to maximize the rate of return on the user's investment.

References:

[1] Quinlan, J. R., Induction of decision trees, *Machine Learning*, Vol.1, 1986, pp. 85-106.

[2] Clark P., and Niblett, T., The CN2 induction algorithm, *Machine Learning*, Vol.3, No.4, 1989, pp. 261-283.

[3] Friedman, N., Geiger, D., and Goldsmidt, M., Bayesian network classifiers, *Machine Learning*, Vol.29, No.2, 1997, pp.131-163.

[4] Vovk, V., Gammerman, A., and Shafer, G., *Algorithmic learning in a random world*, New York: Springer, 2005.

[5] Yu, B., and Zhu, D.-H., Combining neural networks and semantic feature space for email classification, *Knowledge-Based Systems*, Vol.22, 2009, pp. 376-381.

[6] Vapnik V.N., *Statistical learning theory*, New York: Wiley, 1998.

[7] Pai, P.-F., Hsu, M.-F., and Wang, M.-C. , A support vector machine-based model for detecting top management fraud, *Knowledge-Based Systems* in Press, DOI: 10.1016/j.knosys.2010.10.003.

[8] Pawlak, Z., Rough sets, *International Journal of Information and Computer Sciences*, Vol.11, No.5, 1982, pp. 341-356.

[9] Huang, K.Y., and Jane, C.-J., A Hybrid Model for Stock Market Forecasting and Portfolio Selection Based on ARX, Grey System and RS Theories, *Expert Systems With Applications*, Vol.36, 2009, pp. 5387-5392.

[10] Huang, K.Y., A Hybrid GRA / MV Model for the Automatic Selection of Investment Portfolios with Minimum Risk and Maximum Return, *The Journal of Grey System* (ISSN: 0957-3720), Vol.21, 2009, pp. 149-166.

[11] Huang, K.Y., Application of VPRS model with enhanced threshold parameter selection mechanism to automatic stock market forecasting and portfolio selection, *Expert Systems With Applications*, Vol.36, 2009, pp.11652-11661.

[12] Ziarko, W., Variable precision rough set model, *Journal of Computer and System Sciences*, Vol.46, 1993, pp. 39-59.

[13] Ziarko, W., Probabilistic decision tables in the variable precision rough set model, *Computational Intelligence*, Vol.17, 2001, pp. 593-603.

[14] Huang, K.Y., Chang, T.-H., and Chang, T.-C., Determination of the Threshold Value β of Variable Precision Rough Set by Fuzzy Algorithms, *International Journal of Approximate Reasoning* in Press, doi:10.1016/j.ijar.2011.05.001.

[15] Pakhira, M.K., Bandyopadhyay, S., and Maulik, U., Validity index for crisp and fuzzy clusters, *Pattern Recognition*, Vol.37, 2004, pp. 487-501.

[16] Hua, Y.-C., Chena, R.-S., and Tzengb, G.-H., Discovering fuzzy association rules using fuzzy partition methods, *Knowledge-Based Systems*, Vol.16,2003, pp. 137-147.

[17] Bezdek, J.C., Cluster validity with fuzzy sets, *J. Cybernet.*, Vol.3, 1974, pp. 58-74.

[18] Gath, I., and Geva, A. B., Unsupervised optimal fuzzy clustering, *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol.11, 1989, pp.773-781.

[19] Pakhira, M.K., Bandyopadhyay, S., and Maulik, U., A study of some fuzzy cluster validity indices, genetic clustering and application to pixel classification, *Fuzzy Sets and Systems*, Vol.155, 2005, pp. 191-214.

[20] Huang, K.Y., 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*, Vol.37, 2010, pp. 8757-8769.

[21] Clerc, M., and Kennedy, J., The particle swarm-explosion, stability and convergence in a multidimensional complex space, *IEEE Transactions on Evolutionary Computation*, Vol.6, 2002, pp. 58-73.

[22] Lu, H., Pi, E., Peng, Q., Wang, L., and Zhang, C., A particle swarm optimization-aided fuzzy cloud classifier applied for plant numerical taxonomy based on attribute similarity, *Expert Systems with Applications*, Vol.36, 2009, pp. 9388-9397.

[23] Tang, X., Zhuang, L., Cai, J., and Li, C., Multi-fault classification based on support vector machine trained by chaos particle swarm optimization, *Knowledge-Based Systems*, Vol.23, No.5, 2010, pp. 486-490.

[24] Alatas, B., and Akin, E., Multi-objective rule mining using a chaotic particle swarm optimization algorithm, *Knowledge-Based Systems*, Vol.22, No.6, 2009, pp. 455-460.

[25] Huang, K.Y., A Hybrid Particle Swarm Optimization Approach for Clustering and Classification of Datasets, *Knowledge-Based Systems*, Vol.24, No.3, 2009, pp. 420-426.

[26] Clerc, M., The swarm and the queen: towards a deterministic and adaptive particle swarm optimization, in: *Proceedings of the Congress of Evolutionary Computation*, Washington, 1995, pp. 1951-1957.

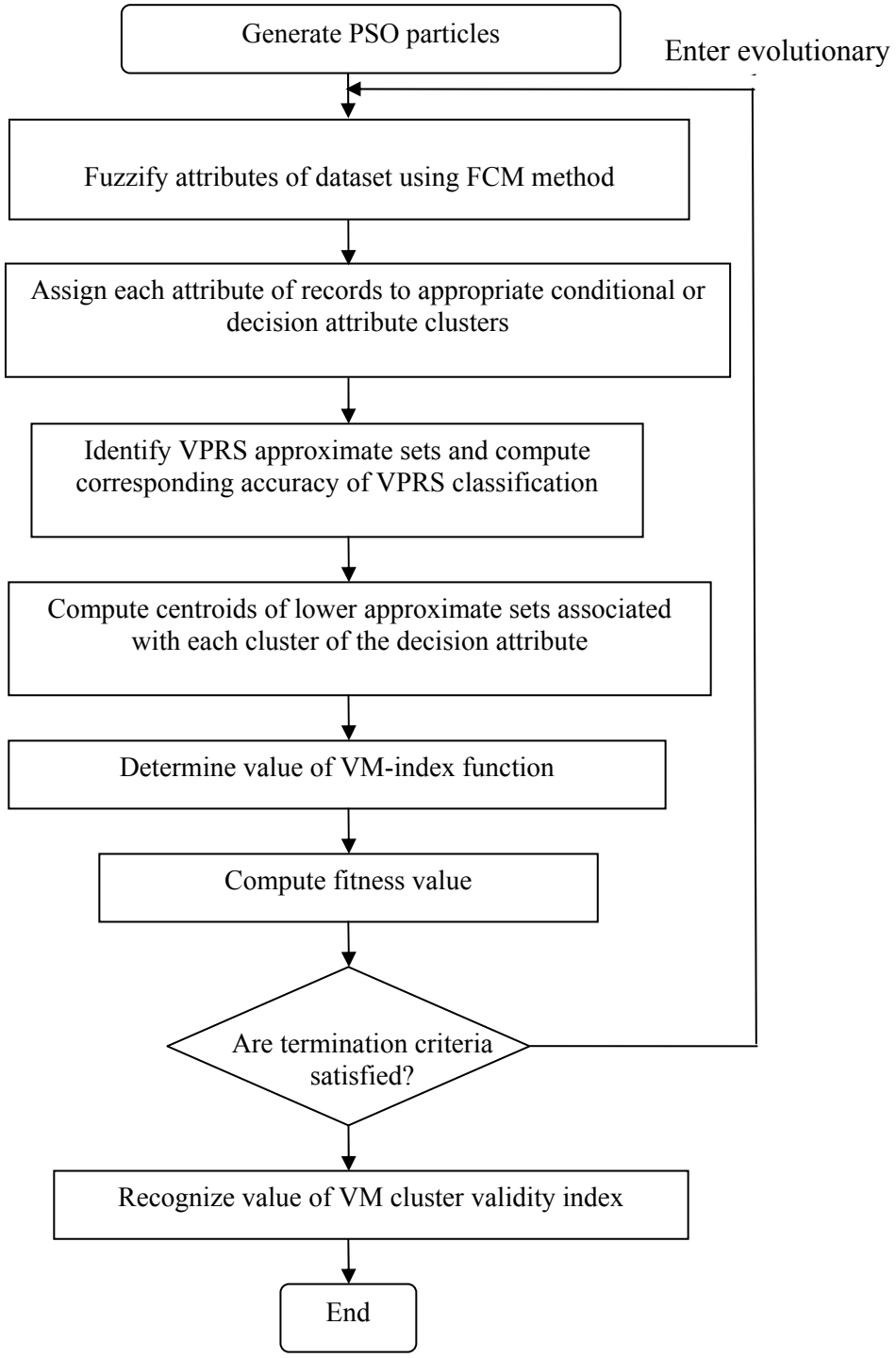

Fig. 1 Flow chart showing basic steps in proposed PSOVPRS-index method.

Table 1 Detailed definitions of VM-index and Huang-based index methods.

Formulation	VM-index	Huang-index [20]
	$VM(C_{M},{}_{\beta}\alpha_{c})=(\frac{1}{C_{M}}\times\frac{\overline{E_{1}}}{{}_{\beta}F'_{N_d}}\times D'_{N_d})$	$H(N_{d},\alpha_{C})=\left(\frac{1}{N_{d}}\times\frac{\overline{E_{1}}}{F'_{N_d}}\times D'_{N_d}\right)$
How to cluster the data	based on a distinct number of clusters to each attribute	based on a constant number of clusters to each attribute
	C_M is the arithmetic mean of numbers of clusters of individual attributes; N_d is the number of clusters of the conditional and decision attributes	N_d is the number of clusters of the conditional and decision attributes
	${}_{\beta}F'_{N_d}=\sum_{c=1}^{N_d}{}_{\beta}E'_{c}$, ${}_{\beta}E'_{c}=\sum_{j=1}^{n}\mu_{cj}^{m'}(x_j(d))\left\|x_j-z'_c\right\|\Big/{}_{\beta}\alpha_c$	$F'_{N_d}=\sum_{c=1}^{N_d}E'_{c}$, $E'_{c}=\sum_{j=1}^{n}\mu_{cj}^{m'}(x_j(d))\left\|x_j-z'_c\right\|\Big/\alpha_c$
	(1) $\mu_{cj}(x_j(d))$ is the membership function of record x_j in the c th cluster of the decision attribute d. (2) z'_c is the multi-dimensional centroid of the lower approximate sets associated with the c -th cluster of the decision attribute d and is obtained by computing the mean values of the conditional and decision attributes of each record within the corresponding sets. (3.1) $\left\|x_j-z'_c\right\|$ is the length of the vector (norm) between the x_j record and z'_c.	
	(3.2) ${}_{\beta}E'_{c}=\sum_{j=1}^{n}\left\|x'_{jc}\right\|\Big/{}_{\beta}\alpha_c$, where $\left\|x'_{jc}\right\|=\mu_{cj}^{m'}(x_j(d))\left\|x_j-z'_c\right\|$; ${}_{\beta}\alpha_c$ is the accuracy of VPRS classification when evaluated in terms of the c -th cluster of the decision attribute.	(3.2) $E'_{c}=\sum_{j=1}^{n}\left\|x'_{jc}\right\|\Big/\alpha_c$, where $\left\|x'_{jc}\right\|=\mu_{cj}^{m'}\left(x_j(d)\right)\left\|x_j-z'_c\right\|$; α_c is the corresponding accuracy of approximation when evaluated in terms of the c th cluster of the decision attribute d
$D'_{N_d}=\max_{i,j=1}^{N_d}\left\|z'_i-z'_j\right\|$ is the maximum separation distance among the centroids of all the lower approximate sets associated with the different clusters of the decision attribute.		

Table 2(a) Membership function values of each attribute of each instance

Code of instances	Conditional attributes				Decision attribute	
	a_1		a_2		d	
1	0.025	0.975	0.061	0.939	0.010	0.990
2	0.063	0.937	0.025	0.975	0.015	0.985
3	0.988	0.012	0.939	0.061	0.985	0.015
4	0.992	0.008	0.974	0.026	0.990	0.010

Table 2(b). Lower approximate sets and upper approximate sets associated with c-th cluster of decision attribute

Code of instances	lower approximate sets $\underline{R}(X : C_d(\mathrm{x}) = c, x \in X)$			
1	2	2	2	$\underline{R}(X : C_d(\mathrm{x}) = 2, x \in X)$
2	2	2	2	
3	1	1	1	$\underline{R}(X : C_d(\mathrm{x}) = 1, x \in X)$
4	1	1	1	

Each of the lower approximate sets $\underline{R}(X : C_D(\mathrm{x}) = c, x \in X)$ is equal to the corresponding upper approximate set $\overline{R}(X : C_D(x) = c, x \in X)$.

Table 2(c) Values of $\left\|x'_{jc}\right\| (= \mu^2_{cj}(x_j(d)) \times \left\|x_j - z'_c\right\|)$

x_j	z'_c	
j	c=1	c=2
1	0.120	0.000
2	0.000	0.000
3	0.615	0.084
4	0.786	0.085
$\sum_{j=1}^{4} \left\|x'_{jc}\right\|$	0.239	0.169

Table 3 Illustrative financial data extracted from TEJ database for first quarter in 2006

Code of companies	(a)	(b)	(c)	(d)	(e)	(f)	(g)
1	-23.4	31.22	54.62	0.07	-0. 34	0.57	-0.03
2	4.14	7.86	3.72	0.231	1.23	2.34	-0.58
3	0.75	0.11	-0.65	0.01	0.73	0.11	2.01
…	…	…	…	…	…	…	…
305	10.22	11.99	1.77	1.50	12.79	15.01	4.01
306	5.27	7.54	2.26	0.94	6.50	9.28	5.17
307	2.38	-1.42	-3.79	-0.18	3.04	-1.81	-2.88

The attributes of columns are (a) Business Profit Rate (b) Pre-Tax Income % (c) Net Non-op.Inc./Rev. (d) PS-Pre_Tax Income (e) Oper.Income/Capital (f) Pre Tax Income/Capital (g) EPS-Net Income

Table 4 Comparison of classification accuracy (CA) obtained from PSOVPRS-index method, PSORS-index method, and pseudo-supervised decision-tree classification method for 10-fold subsamples.

	PSOVPRS-index method		PSORS-index method		Pseudo-supervised decision-tree classification method	
i th subsamples	training dataset	testing dataset	training dataset	testing dataset	training dataset	testing dataset
1	0.76	1.00	0.69	0.94	0.34	0.13
2	0.85	1.00	0.81	1.00	0.31	0.06
3	0.91	1.00	0.76	1.00	0.30	0.16
4	0.80	1.00	0.72	1.00	0.27	0.16
5	0.77	1.00	0.68	1.00	0.30	0.03
6	0.74	1.00	0.71	1.00	0.25	0.10
7	0.77	0.87	0.72	0.87	0.29	0.13
8	0.81	1.00	0.77	0.93	0.26	0.07
9	0.77	1.00	0.76	0.93	0.30	0.27
10	0.78	1.00	0.77	1.00	0.31	0.17
average CA	0.79	0.99	0.74	0.97	0.29	0.13
deviation of CA	4.87%	4.08%	4.11%	4.60%	2.75%	6.74%