Pattern Recognition 44 (2011) 1738–1749

Contents lists available at ScienceDirect

Pattern Recognition

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

Shadowed sets in the characterization of rough-fuzzy clustering

Jie Zhou [a,b,*], Witold Pedrycz [b,c], Duoqian Miao [a]

[a] Department of Computer Science and Technology, Tongji University, Shanghai 201804, PR China
[b] Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada T6G 2G7
[c] System Research Institute, Polish Academy of Sciences, Warsaw, Poland

ARTICLE INFO

Article history:
Received 5 August 2010
Received in revised form
17 January 2011
Accepted 21 January 2011
Available online 27 January 2011

Keywords:
Shadowed sets
Rough sets
Rough-fuzzy clustering
Granulation–degranulation

ABSTRACT

In this study, we develop a technique of an automatic selection of a threshold parameter, which determines approximation regions in rough set-based clustering. The proposed approach exploits a concept of shadowed sets. All patterns (data) to be clustered are placed into three categories assuming a certain perspective established by an optimization process. As a result, a lack of knowledge about global relationships among objects caused by the individual absolute distance in rough *C*-means clustering or individual membership degree in rough-fuzzy *C*-means clustering can be circumvented. Subsequently, relative approximation regions of each cluster are detected and described. By integrating several technologies of Granular Computing including fuzzy sets, rough sets, and shadowed sets, we show that the resulting characterization leads to an efficient description of information granules obtained through the process of clustering including their overlap regions, outliers, and boundary regions. Comparative experimental results reported for synthetic and real-world data illustrate the essence of the proposed idea.

1. Introductory comments

Real-world data distribution often involves ambiguous structures characterized by uncertainty and overlap between elements of the structure (clusters). The main task of clustering is to partition an unlabeled dataset $\{x_1, x_2, \ldots, x_N\}$, each object $x_i \in \mathbb{R}^n$, into C $(1 < C < N)$ subgroups such that the objects in the same cluster are characterized by the highest levels of similarity (homogeneity). During the realization of clustering algorithms, one can highlight several important issues.

K-Means [1] being regarded as a classical prototype (centroid)-based partitive clustering method, assigns each object to exactly one cluster. Though *K*-Means is effective, its usefulness degenerates when dealing with overlapping clusters. Fuzzy clustering, especially Fuzzy *C*-Means (FCM) [2], as the extension of *K*-Means, is often used to reveal the structure of a dataset and to construct information granules. It utilizes a partition matrix to capture the degree of each object belonging to each cluster, so the overlapping circumstances can be effectively described. The main challenge to FCM is the sensitivity to noisy objects.

Recently, considering rough set theory [3], Lingras and West [4] introduced Rough *C*-Means (RCM) clustering, which describes each cluster not only by a prototype, but also with a pair of lower and upper bounds (interval set). Weighted parameters are used to measure the importance of lower bounds and boundary regions when calculating new prototypes. RCM can deal with the uncertainty and vagueness arising in the boundary region of each cluster. Since no memberships are involved, the closeness of objects to the clusters cannot be detected [5].

As two important paradigms of Granular Computing [6,7], rough sets and fuzzy sets have been developed separately to a significant extent. However, they are also complementary. Involving membership degrees, Mitra et al. [5] put forward a Rough-Fuzzy *C*-Means (RFCM) clustering method, which integrates the advantages of the technologies of fuzzy sets and rough sets. The lower and upper bounds are determined according to the membership degrees, not the individual absolute distances between an object and its neighbors. Maji et al. [8] further pointed out that the objects in the lower bound of a cluster should have similar influence on this cluster and the corresponding prototype, and their weights should also be independent of other prototypes when iteratively computing the new prototypes. Following this notion, Maji modified the computation for new prototypes under the scheme of the RFCM.

No matter which rough set-based partitive clustering methods will be used, their pertinent parameters have to be carefully optimized. One of them is the threshold that determines the

* Corresponding author at: Department of Computer Science and Technology, Tongji University, Shanghai 201804, PR China. Tel.: +86 15000600177; fax: +86 21 69589979.
E-mail addresses: jie_jpu@163.com (J. Zhou), pedrycz@ee.ualberta.ca (W. Pedrycz), miaoduoqian@163.com (D. Miao).

0031-3203/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2011.01.014

approximation regions for each cluster. The other is the weighted measures evaluating the importance of lower bounds and boundary regions when updating the prototypes in iterations. Though the initial configuration of the methods can be optimized by a genetic algorithm [9], the selection of parameters mainly depends on subjective tuning in some available research and the obtained results need more interpretations [4,11]. In addition, since only the individual absolute distance and individual membership degree are, respectively, exploited in the RCM and RFCM, the approximation regions that form the prototypes might be deflected when some outliers are involved [10].

Shadowed sets [12], which are considered as a conceptual and algorithmic bridge between rough sets and fuzzy sets, have become a new emerging paradigm of Granular Computing being successfully used for unsupervised learning, resulting in a so-called Shadowed *C*-Means (SCM) [13]. Unlike FCM, the weighted values of objects at the core level of a cluster are enhanced in the SCM. The membership degrees of these objects to this cluster should also be uniform when calculating the dfsfsa corresponding prototype, which is the same as in Maji's notion. The weighted values of objects at the exclusion level of a cluster will be reduced by raising the fuzzification coefficient in the form of a double exponential. Compared with the FCM, the capability of SCM when dealing with outliers is enhanced and improved clustering results can be envisioned [13].

In this study, we concentrate on the determination of the threshold parameter in three types of rough set-based clustering methods including RCM, RFCM, and Maji's method. According to the optimization process supported by shadowed sets, this user-defined threshold becomes automatically selected based on the data's intrinsic structural complexity. The lack of knowledge about global relationships among objects caused by the individual absolute distance in RCM or individual membership degree in RFCM can be circumvented. Therefore, comparative accurate approximation regions of each cluster can be detected which are crucial to the calculations of the associated prototype. Furthermore, a new validity index is proposed by taking into account the granulation–degranulation principle and its underlying mechanism. It is worth noting here that this concept is quite different from the idea supported by cluster validity indices available in the literature including such alternatives as PBM [14], DB [15] and XB indices [16].

By integrating various technologies of Granular Computing involving fuzzy sets, rough sets and shadowed sets, some significant merits of the proposed development can be offered. The membership degrees can effectively describe an overlapping effect present in the partition matrices. In particular, the concept of approximation regions can deal with uncertainty and vagueness arising in the boundary region of any cluster, while the shadowed sets make the modified algorithms robust when coping with noisy objects. Experimental results for synthetic and real-world data show the comparative performance of the proposed notion with respect to the new index along with other available validity indices.

The structure of the paper is as follows. Some basic concepts of rough sets are briefly introduced in Section 2. Section 3 reviews the pertinent rough set-based clustering methods along with their generalized version. In Section 4, we provide shadowed sets as a vehicle for describing information granules obtained through the process of clustering. Based on granulation–degranulation mechanisms, a new cluster validity index is presented in Section 5. Section 6 includes the results of experiments involving both synthetic and real datasets. In Section 7, main conclusions are covered.

Throughout the study, we adhere to the following notation:

N	number of objects;
C	number of clusters;
U_i	ith cluster;
$\mathbf{v}_i$	ith prototype;
$\mathbf{x}_k$	kth object;
u_{ik}	membership of $\mathbf{x}_k$ in U_i;
m	fuzzification coefficient;
$\underline{R}U_i$	lower bound of U_i;
$\overline{R}U_i$	upper bound of U_i;
R_bU_i	boundary region of U_i;
δ_j	standard deviation of the jth feature;
$d(\mathbf{x}_k, \mathbf{v}_i)$	distance between $\mathbf{x}_k$ and $\mathbf{v}_i$;
$card(X)$	cardinality of set X.

2. A brief review of rough sets

Rough sets aim at forming an approximate definition for a target set in terms of some definable sets, especially, when the target set is uncertain or imprecise. Some basic concepts in the rough set theory are briefly recalled in this section. More detailed discussion can be found in [3,17].

Let U denote a finite nonempty universe. A is a set of features (attributes) that describe the objects in the universe. A can be defined as an equivalence relation, referred to as an indiscernibility relation on U, with which U can be partitioned into a collection of disjoint equivalence classes $U/A=\{E_1,E_2,\ldots,E_{card(U/A)}\}$. $card(X)$ stands for the cardinality of set X. Each $E_i \in U/A$ is called an elementary set. Any arbitrary subset (target set) $X \subseteq U$ can be represented in terms of a pair of upper and lower bounds $\overline{A}X$ and $\underline{A}X$ which are defined as follows:

$$\overline{A}X = \cup\{E|E\cap X\neq\emptyset,\ E\in U/A\},\quad \underline{A}X = \cup\{E|E\subseteq X,\ E\in U/A\}. \tag{1}$$

The upper bound $\overline{A}X$ is composed of objects that have a nonempty intersection with X, namely belong to the set X possibly. The lower bound $\underline{A}X$ is composed of objects that are subsets of X, namely belong to the set X certainly. $U-\overline{A}X$ is called the negative region of X, in which the objects do not belong to the set X. The objects positioned in-between the lower and upper bounds form the boundary region of X. If the boundary region is empty, X is called a crisp set. Otherwise, we are concerned with a rough set. The upper and lower bounds approximate the set X from two sides. In other words, X can be approximately represented by two sets. If the target set X is uncertain or vague, such approximate descriptions have an important meaning.

3. Rough set-based partitive clustering

In this section, some rough set-based partitive clustering algorithms will be revisited which include rough *C*-means algorithm (Lingras' model) and two types of rough-fuzzy *C*-means algorithms (Mitra's model and Maji's model).

3.1. Rough C-means

Lingras et al. [4] extended the concept of rough approximations to develop a clustering algorithm in which the following basic rough set properties need to be satisfied.

Property 1. *An object can belong to the lower bound of one cluster at most.*

Property 2. *An object that belongs to the lower bound of a cluster also belongs to the upper bound of this cluster.*

Property 3. *An object that does not belong to any lower bound will belong to more than one upper bound.*

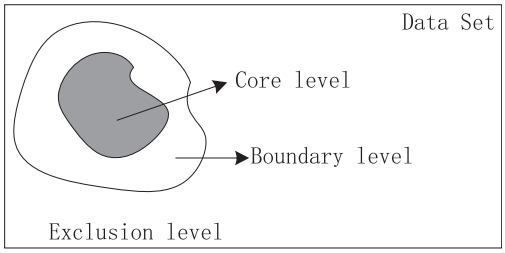

Fig. 1. Three levels of belongingness with respect to a fixed cluster.

Each cluster has its own lower and upper bounds. The new prototype calculations will only depend upon the objects in these two approximation regions, not all objects as in the *K*-Means, FCM or SCM. Thus the useless information can be filtered out and ensuing numeric computing can be reduced. For a fixed cluster, all objects are split into three categories, namely, core level, boundary level and exclusion level, as shown in Fig. 1.

The objects located at the core level definitely belong to this cluster. The objects at the boundary level possibly belong to this cluster, viz., they come with some component of vagueness and uncertainty. Other objects that fall within the exclusion level do not belong to this cluster. The contributions of objects located at different levels to the cluster are distinct. Generally, the objects present at the core level exhibit the highest importance while the objects positioned in the exclusion region are almost ignored.

Suppose N objects are grouped into C clusters $U_1,U_2,\ldots,U_C$. The corresponding prototypes $\mathbf{v}_1,\mathbf{v}_2,\ldots,\mathbf{v}_C$, $\mathbf{v}_i \in \mathbb{R}^n$, are updated in the following way.

$$\mathbf{v}_i = \begin{cases} w_l A_1 + w_b B_1 & \text{if } \underline{R}U_i \neq \emptyset \wedge R_b U_i \neq \emptyset, \\ B_1 & \text{if } \underline{R}U_i = \emptyset \wedge R_b U_i \neq \emptyset, \\ A_1 & \text{if } \underline{R}U_i \neq \emptyset \wedge R_b U_i = \emptyset, \end{cases} \tag{2}$$

where

$$A_1 = \frac{\sum_{\mathbf{x}_k \in \underline{R}U_i} \mathbf{x}_k}{card(\underline{R}U_i)}, \tag{3}$$

$$B_1 = \frac{\sum_{\mathbf{x}_k \in R_b U_i} \mathbf{x}_k}{card(R_b U_i)}. \tag{4}$$

$R_bU_i = \overline{R}U_i - \underline{R}U_i$ denotes the boundary region of cluster U_i, where $\underline{R}U_i$ and $\overline{R}U_i$ denote the lower and upper bounds of cluster U_i with respect to feature set R, respectively. A_1, B_1 can be considered as the contributions by the lower bounds and boundary regions, respectively. $w_l (0.5 < w_l \leq 1)$ and $w_b = 1 - w_l$ are the weights for these two contributed parts. When updating a prototype, the higher the value of w_l, the more important the lower bound is. There is no need to consider the cases that both the lower bound and boundary region of a cluster are empty since this cluster has no representative [10].

In order to determine the lower bound and boundary region of each cluster, Lingras et al. [4] utilized the following rules:

If $d(\mathbf{x}_k,\mathbf{v}_q) - d(\mathbf{x}_k,\mathbf{v}_p) \leq \varepsilon$, then $\mathbf{x}_k \in \overline{R}U_p$ and $\mathbf{x}_k \in \overline{R}U_q$. In this case, $\mathbf{x}_k$ cannot belong to the lower bound of any cluster. Otherwise, $\mathbf{x}_k \in \underline{R}U_p$. Here $d(\mathbf{x}_k, \mathbf{v}_i)$ denotes the distance between object $\mathbf{x}_k$ and prototype $\mathbf{v}_i$ $(i=1,2,\ldots,C)$. $d(\mathbf{x}_k, \mathbf{v}_p)$ and $d(\mathbf{x}_k, \mathbf{v}_q)$ stand for the minimum and secondary minimum of $\mathbf{x}_k$ over all clusters, respectively. A weighted Euclidean distance will be used in this study, which is expressed as follows:

$$d(\mathbf{x}_k,\mathbf{v}_i) = \sqrt{\sum_{j=1}^{n} \frac{(x_{kj} - v_{ij})^2}{\delta_j^2}}, \tag{5}$$

where δ_j is the standard deviation of the jth feature. Compared with the standard Euclidean distance, its weighted version eliminates the influence of significantly different ranges of individual features.

The threshold ε is crucial for the determination of the approximation regions of each cluster. The lower the threshold value, the more objects will belong to the lower bounds. To the contrary, the higher the threshold, the more objects will belong to the boundary regions. The improperly selected value of the threshold will result in inaccurate approximation regions, which then misguide the formation of the prototypes. In addition, since no membership degrees are involved, the overlapping partitions cannot be effectively handled by the RCM.

3.2. Rough-fuzzy C-means

Incorporating fuzzy clustering methods, Mitra et al. [5] put forward the version of Rough-Fuzzy *C*-means (referred to as RFCM I) in which membership degree u_{ik} will replace the absolute distance $d(\mathbf{x}_k, \mathbf{v}_i)$ when determining the approximation regions for each cluster. This adjustment will enhance the robustness of the clustering when dealing with overlapping situations. In this case, the calculation of prototypes is governed by the following expressions:

$$\mathbf{v}_i = \begin{cases} w_l A_2 + w_b B_2 & \text{if } \underline{R}U_i \neq \emptyset \wedge R_b U_i \neq \emptyset, \\ B_2 & \text{if } \underline{R}U_i = \emptyset \wedge R_b U_i \neq \emptyset, \\ A_2 & \text{if } \underline{R}U_i \neq \emptyset \wedge R_b U_i = \emptyset, \end{cases} \tag{6}$$

where

$$A_2 = \frac{\sum_{\mathbf{x}_k \in \underline{R}U_i} u_{ik}^m \mathbf{x}_k}{\sum_{\mathbf{x}_k \in \underline{R}U_i} u_{ik}^m}, \tag{7}$$

$$B_2 = \frac{\sum_{\mathbf{x}_k \in R_b U_i} u_{ik}^m \mathbf{x}_k}{\sum_{\mathbf{x}_k \in R_b U_i} u_{ik}^m}. \tag{8}$$

A_2 and B_2 can be considered as the contributors to the *fuzzy* lower bounds and *fuzzy* boundary regions, respectively. As in the RCM, the weights $w_b = 1 - w_l$ and $0.5 < w_l \leq 1$. In order to determine the approximation regions, the following calculations are completed.

If $u_{pk} - u_{qk} \leq \varepsilon$, then $\mathbf{x}_k \in \overline{R}U_p$ and $\mathbf{x}_k \in \overline{R}U_q$. In this case, $\mathbf{x}_k$ cannot belong to the lower bound of any cluster. Otherwise, $\mathbf{x}_k \in \underline{R}U_p$. u_{ik} denotes the membership degree of object $\mathbf{x}_k$ to the cluster with prototype $\mathbf{v}_i$ $(i=1,2,\ldots,C)$ and is calculated in the same way as realized in the FCM. u_{pk} and u_{qk} represent the maximum and secondary maximum of $\mathbf{x}_k$ over all clusters, respectively.

The fuzzification coefficient m assumes values greater than 1. Its value reflects the geometry of fuzzy clusters [18]. When the value is close to 1, it implies a Boolean nature of the cluster. On the other hand, it will result in spike-like membership functions when the value increases (such as three or more). By choosing different values of m, we can control the shape of clusters. Yu et al. [21] provided a theoretical basis for selecting the fuzzification coefficient and pointed out that its suitable values should depend on the dataset itself. A fuzzy encoding and decoding mechanism [22] has also been constructed for choosing experimental optimal values. Predominantly, applications involving FCM often set this value to be equal to 2.

Maji et al. [8] pointed out that the weights of objects forming the lower bound of a cluster should be independent of other prototypes and they should have the same contribution to this cluster. Nevertheless, the objects in the boundary region should exhibit different influence on this prototype. Following these observations, Mitra's model is modified where the prototypes are computed depending on the weighted average of the lower

bounds and *fuzzy* boundary regions. More specifically, we have

$$\mathbf{v}_i = \begin{cases} w_l A_1 + w_b B_2 & \text{if } \underline{R}U_i \neq \emptyset \wedge R_b U_i \neq \emptyset, \\ B_2 & \text{if } \underline{R}U_i = \emptyset \wedge R_b U_i \neq \emptyset, \\ A_1 & \text{if } \underline{R}U_i \neq \emptyset \wedge R_b U_i = \emptyset. \end{cases} \quad (9)$$

The parameters w_l and ε as well as the rules used to determine the approximation regions are the same as encountered in the RFCM I. It has been shown that the performance of the modified RFCM (referred here to as RFCM II) is better than RFCM I according to some proposed rough set-based quantitative indices [8].

3.3. Generalized rough set-based C-means algorithm

According to the common properties of RCM, RFCM I and RFCM II, a generalized version of rough set-based *C*-means algorithm can be described as follows:

Algorithm 1. Generalized rough set-based *C*-means algorithm

Step 1: Initialization. Assign initial prototypes for the *C* clusters;

Step 2: Determine the lower bound and boundary region of each cluster;

Step 3: Update the prototypes for the *C* clusters;

Step 4: Repeat Steps 2 and 3 until convergence has been reached.

Convergence pointed to in Step 4 means the obtained prototypes in the current iteration are identical to those that have been generated in the previous one, namely, the prototypes are stabilized. Steps 2 and 3 are the main points in the generalized version. The modifications based on original RCM, namely, RFCM I and RFCM II, are all concentrated on them. Compared with *K*-Means and FCM, the objects are divided into three regions with respect to a given cluster. The contributions for the prototype and cluster from the lower bound (core level) will be enhanced and the contributions from the boundary region will be diminished relative to the contributions encountered in the FCM.

The accurate approximation regions and reasonable values of weights directly affect the clustering results. However, a single threshold cannot reflect the differences among all clusters and the closeness of objects to the clusters will not be effectively described. In this case, approximation regions may be distorted and the prototypes may deviate from their expected locations. In order to form accurate regions, we anticipate that each cluster should come with a suitable threshold reflecting structural characteristics of the data when being perceived from the perspective of some structural relationships.

4. Shadowed set-based rough-fuzzy clustering

Shadowed set-based rough-fuzzy clustering methods are proposed in this section. We show that the threshold parameter that affects the lower bound and boundary region of each cluster can be decided upon automatically. Its value can be adjusted according to the structure of data and its complexity.

4.1. Shadowed sets

Shadowed sets, as introduced by Pedrycz [12], is one among several key contributions to the area of Granular Computing. It could be considered as new and stand-alone constructs, yet it is often induced by the corresponding fuzzy sets. It is simpler and more practical than fuzzy sets and can be sought as a symbolic representation of numeric fuzzy sets [19].

Three quantification levels being elements of the set {0, 1, [0, 1]} are utilized to simplify the relevant fuzzy sets in shadowed set theory. Obviously, it not only simplifies the interpretation but also avoids a number of computations of numeric membership grades comparing with the methodology of fuzzy sets. Conceptually, shadowed sets are close to rough sets even though their mathematical foundations are very different. The concepts of negative region, lower bound and boundary region in rough set theory are corresponding to three-logical values 0, 1, and [0,1] in shadowed sets, namely, excluded, included and uncertain, respectively. In this sense, shadowed sets can be considered as the bridge between fuzzy and rough sets.

The construction of shadowed sets is based on balancing the uncertainty that is inherently associated with fuzzy sets, in other words, uncertainty relocation. As elevating membership values (high enough) of some regions of universe to 1 and at the same time, reducing membership values (low enough) of some regions of universe to 0, we can eliminate the uncertainty in these regions. In order to balance the total uncertainty, it needs to compensate these changes by allowing for the emergence of uncertainty regions, namely, it results in shadowed sets.

Given a continuous fuzzy membership function $x \to f(x)$, $f(x) \in [0,1]$, the reduction of uncertainty and shadows can be represented as in Fig. 2 and are quantified as follows.

Reduction of membership:

$$\Omega_1 = \int_{x:f(x) \leq \alpha} f(x)\, dx. \quad (10)$$

Elevation of membership:

$$\Omega_2 = \int_{x:f(x) \geq 1-\alpha} (1-f(x))\, dx. \quad (11)$$

Formation of shadows:

$$\Omega_3 = \int_{x:\alpha < f(x) < 1-\alpha} dx. \quad (12)$$

The separate threshold α in shadowed sets can be optimized by realizing the principle of uncertainty balance. It translates into the minimization of the following objective function.

$$V(\alpha) = |\Omega_1 + \Omega_2 - \Omega_3|. \quad (13)$$

The optimal threshold α satisfies the requirement $\alpha_{opt} = \text{argmin}_\alpha V(\alpha)$, where $\alpha \in [0,0.5)$. The discrete version of optimization process can be expressed in a similar manner. Suppose $u_1, u_2, \ldots, u_N$ are discrete membership values, $u_k \in [0,1]$ $(k=1,2,\ldots,N)$. $u_{\max}$ and $u_{\min}$ denote the maximal and minimal values, respectively. The objective function is modified as

$$V(\alpha) = |\psi_1 + \psi_2 - \psi_3|, \quad (14)$$

where $\psi_1 = \sum_{u_i \leq \alpha} u_i$ means the reduction of membership. $\psi_2 = \sum_{u_i \geq (u_{\max}-\alpha)} (u_{\max} - u_i)$ means the elevation of membership. $\psi_3 = card(\Delta)$ represents the shadows, $\Delta = \{i | \alpha < u_i < (u_{\max} - \alpha)\}$. The range of feasible values of threshold α is suggested in $[u_{\min}, (u_{\min} + u_{\max})/2]$.

Three logical values induced by shadowed sets correspond to the notions of three approximation regions in rough set theory. Though the foundations of these two methodologies are different, they share some common philosophies when coping with uncertain

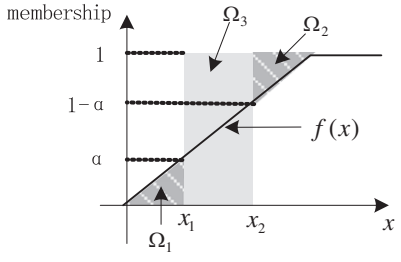

Fig. 2. Shadowed sets induced by fuzzy membership function $f(x)$.

problems. The main merits of shadowed sets involve the optimization mechanism for choosing separate threshold and the reduction of the burden of plain numeric computations.

4.2. Shadowed set-based rough-fuzzy clustering

The membership degrees of objects belonging to a fixed cluster U_i ($i=1,2,\ldots,C$) can be considered as a generic fuzzy set. Under this consideration, we can determine the approximation regions for cluster U_i by integrating shadowed sets. The algorithm is described as follows.

Algorithm 2. Determine the approximation regions based on shadowed sets

Step 1: Compute membership values u_{ik} of each object $\mathbf{x}_k$ to each prototype $\mathbf{v}_i$,

$$u_{ik}=\frac{1}{\sum_{j=1}^{C}\left(\frac{d(\mathbf{x}_k,\mathbf{v}_i)}{d(\mathbf{x}_k,\mathbf{v}_j)}\right)^{2/(m-1)}}; \tag{15}$$

*Step*2: Based on the optimization process in shadowed sets, compute optimal threshold α_i for each cluster U_i,

$$\alpha_i=\operatorname*{argmin}_{\alpha}(V_i),$$

where

$$V_i=\left|\sum_{k:u_{ik}\le\alpha}u_{ik}+\sum_{k:u_{ik}\ge(\max_k(u_{ik})-\alpha)}\left(\max_k(u_{ik})-u_{ik}\right)-card\left(\left\{k|\alpha<u_{ik}<\left(\max_k(u_{ik})-\alpha\right)\right\}\right)\right|; \tag{16}$$

*Step*3: According to α_i, determine the lower bound and boundary region of each cluster U_i,

$$\underline{R}U_i=\left\{\mathbf{x}_k|u_{ik}\ge\left(\max_k(u_{ik})-\alpha_i\right)\right\}, \tag{17}$$

$$R_bU_i=\left\{\mathbf{x}_k|\alpha_i<u_{ik}<\left(\max_k(u_{ik})-\alpha_i\right)\right\}. \tag{18}$$

where $u_{ik}\in[0,1]$ ($i=1,2,\ldots,C, k=1,2,\ldots,N$), $\sum_{i=1}^{C}u_{ik}=1$ and $0<\sum_{k=1}^{N}u_{ik}<N$. After Algorithm 2 is completed, each cluster comes with its lower bound and boundary region. Here, the threshold is not subjectively user-defined but it is established on the balance of uncertainty and can be adjusted automatically in the clustering process. In addition, the determination of approximation regions is not dependent on the individual absolute distance or the individual membership value. It considers all membership values with respect to a fixed cluster when updating the prototype of this cluster. Thus the three levels of objects regarding this cluster can be effectively divided.

Based on Algorithm 2, the generalized version of rough set-based clustering algorithm can be refined as follows.

Algorithm 3. Shadowed set-based rough-fuzzy clustering

Step 1: Assign a random membership partition matrix $\{u_{ik}\}$;

Step 2: Based on shadowed sets, compute optimal α_i for each cluster U_i ($i=1,2,\ldots,C$);

Step 3: According to α_i, determine the lower bound and boundary region for each cluster U_i;

Step 4: Calculate the prototypes by formula (2), (6) or (9);

Step 5: Update the membership partition matrix $\{u_{ik}\}$;

Step 6: Repeat Steps 2–5 until convergence has been reached.

Algorithm 3 will be referred to as shadowed set-based rough *C*-means (SRCM), shadowed set-based rough-fuzzy *C*-means I (SRFCM I) and shadowed set-based rough-fuzzy *C*-means II (SRFCM II) according to formulas (2), (6) and (9) used in Step 4, respectively.

The main difference between Algorithm 3 and available rough set-based clustering methods is the mechanism for choosing a suitable threshold for each cluster. The threshold values used in the RCM and RFCM are often user-defined and the approximation regions are determined from the perspective of individual objects, then the global knowledge over all objects when calculating the prototype for each cluster will be lost. However, the threshold in Algorithm 3 will be automatically adjusted and optimized. The approximation regions are determined from the perspective of individual clusters and the accurate three levels can be availably detected. In addition, the membership computation can capture the overlapping partitions and the concept of approximation regions can handle the uncertain arising from the boundary regions. By integrating fuzzy sets, rough sets and shadowed sets, the proposed notion can effectively deal with uncharted situations.

Like most partitive clustering methods, the shadowed set- and rough set-based clustering approaches cannot effectively cope with non-sphere datasets. In this case, more information about the data structure is expected to be integrated.

5. A validity index based on granulation–degranulation mechanisms

During the recent years, some validity indices are proposed to evaluate clustering methods which include fuzzy and non-fuzzy versions, such as PBM, DB and XB indices. These indices often follow the principle that the distance between objects in the same cluster should be as small as possible and the distance between objects in different clusters should be as large as possible. They have also been used to acquire the optimal number of clusters *C* [14]. However, each of them can work better than others depending on the selected datasets [20]. In what follows, we introduce a new validity index, which is based on the granulation–degranulation mechanisms that are schematically shown in Fig. 3.

Essentially, fuzzy clustering process can be treated as a granulation mechanism. Then the information granules established here are expected to reflect the original data as much as possible, so that the input objects should be represented in terms of information granules, involving prototypes and associated membership degrees. In the subsequent step, degranulation process takes place and is applied to original objects and reconstructed (de-granulated) based on the prototypes and the partition matrix. The results of degranulation are expected to be as close as possible to the original objects subject to the granulation. The concept of granulation–degranulation comes also under the name of fuzzification-defuzzification, coding–decoding, compression–decompression and alike. Recently, the mechanisms of granulation–degranulation are utilized in the design of adjustable fuzzy clustering [22].

Formally, given an object $\mathbf{x}$, suppose that $\hat{\mathbf{x}}$ is the corresponding result of degranulation. An overall performance of the granulation–degranulation mechanisms is quantified as follows:

$$Q=\sum_{k=1}^{N}d^2(\mathbf{x}_k,\hat{\mathbf{x}}_k), \tag{19}$$

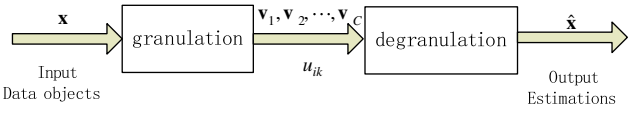

Fig. 3. A schematic view at the granulation–degranulation mechanisms.

where

$$\hat{\mathbf{x}}_k = \frac{\sum_{i=1}^{C} u_{ik}^m \mathbf{v}_i}{\sum_{i=1}^{C} u_{ik}^m}. \quad (20)$$

In general, the values of Q will decrease with the increase of the number of clusters C. The reason is that more clusters can provide the detailed information of data structure, so that the estimation of each object can be precisely established in the degranulation process. The smaller the value of Q, the better the established information granules will be reflected under the same value of C.

6. Experimental studies

We report on the results produced by different clustering algorithms, two synthetic two-dimensional datasets and some datasets coming from the UCI repository [23].

6.1. Synthetic dataset I

This synthetic dataset is a mixture of Gaussian distributions as depicted in Fig. 4. It consists of three clusters with 50 data per cluster. Two of the three clusters exhibit some overlap.

The results obtained by running FCM and including prototypes and the corresponding membership degrees constitute an initial configuration for the implementation of SCM, SRCM, SRFCM I, and SRFCM II. Since the lower bound of each cluster forms the main contribution for this cluster, its weighted value should be relatively higher [5,8]. Here, set w_l=0.95 and m=2. They are kept constant for all datasets and all experiments. In order to calculate the optimal threshold α for each cluster in shadowed set-based methods, its value is varied from $u_{\min}$ to $(u_{\min}+u_{\max})/2$ by small steps equal to 0.001 and the value for which the performance index V_i attains minimum becomes selected as a solution. All the algorithms were run on a personal computer with Intel Pentium Dual-Core T5870 2.0 GHz processor and 1 Gb RAM.

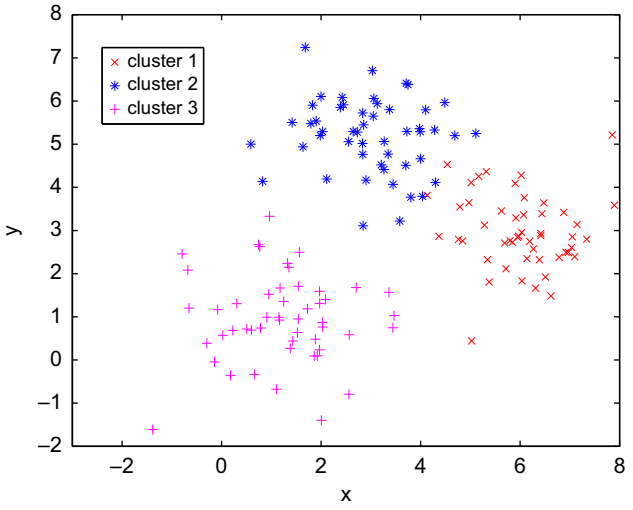

Fig. 4. Scatter-plot of synthetic dataset I.

The prototypes obtained by each method are presented in Table 1. As Fig. 5 shows, each shadowed set-based clustering algorithm can separate well the core level (lower bound) and boundary region of each cluster. Moreover, cluster 2 acquires the same results under different shadowed set-based clustering methods. However, the results in cluster 3 generated by the SRCM, refer to Fig. 5(b), exhibit some minor differences when compared with the results produced by the other three methods, refer to Fig. 5(a, c and d). One data object in the cluster 3 is partitioned to the core level of this cluster by SRCM and is partitioned to the boundary region of this cluster by other three methods. In addition, one data object in the cluster 3 is partitioned to the boundary region of this cluster by SRCM and is partitioned to the core level of this cluster by the others. The results in cluster 1 generated by the SRFCM I, refer to Fig. 5(c), also show some minor differences compared with the results provided by other methods. The obtained approximation regions could be distorted due to some objects that are displaced, even if the number of these objects is very small.

Furthermore, it can be observed that some objects only belong to the boundary region of one cluster, meaning that these objects only belong to the upper bound of one cluster which indicates that the third property in Lingras' model needs not to be always satisfied. The reason behind this effect is that the lower bounds and the boundary regions are being formed from the perspective of each cluster, not the individual objects, and these are independent from any other clusters. In the case of increased overlap between clusters, more objects tend to appear in the common boundary region as seen between the first and the second clusters.

The threshold values that determine the lower bound and the boundary region of each cluster are adjusted automatically according to the intrinsic structural complexities of data detected during the implementation. The obtained threshold values are distinct for different shadowed set-based clustering methods as illustrated in Table 2. According to these obtained threshold values, the lower bound and boundary region of each cluster are depicted in Fig. 5 (right column). It is noticeable that the lower bounds do not intersect which is not the case for some boundary regions of different clusters.

To compare the results obtained by the introduced partitive clustering algorithms, some validity indices are utilized including PBM, XB, DB indices as well as the reconstruction index Q. The obtained results are presented in Table 3. Note that the greater the values of the PBM index and the smaller the values of the XB, DB and Q indices, the better the clustering results are. It becomes apparent that shadowed set-based clustering methods perform far better than the generic FCM. Furthermore, clustering utilizing shadowed sets and rough sets performs better than the SCM method. This implies that the partition of the approximation regions can better capture the existing data structure. Data objects located in different regions (core, boundary and exclusion) exhibit different levels of contribution to prototypes and clusters.

Table 1
Prototypes obtained for the synthetic dataset I.

	FCM		SCM		SRCM		SRFCM I		SRFCM II	
Prototype 1	6.041	2.8952	6.0772	2.9018	6.0726	2.8906	6.1688	2.8569	6.0895	2.8901
Prototype 2	2.9681	5.251	2.9488	5.3166	2.9416	5.3262	2.919	5.3604	2.9407	5.3317
Prototype 3	1.1981	0.89709	1.1174	0.82937	1.17	0.78534	1.1259	0.80343	1.0952	0.8046

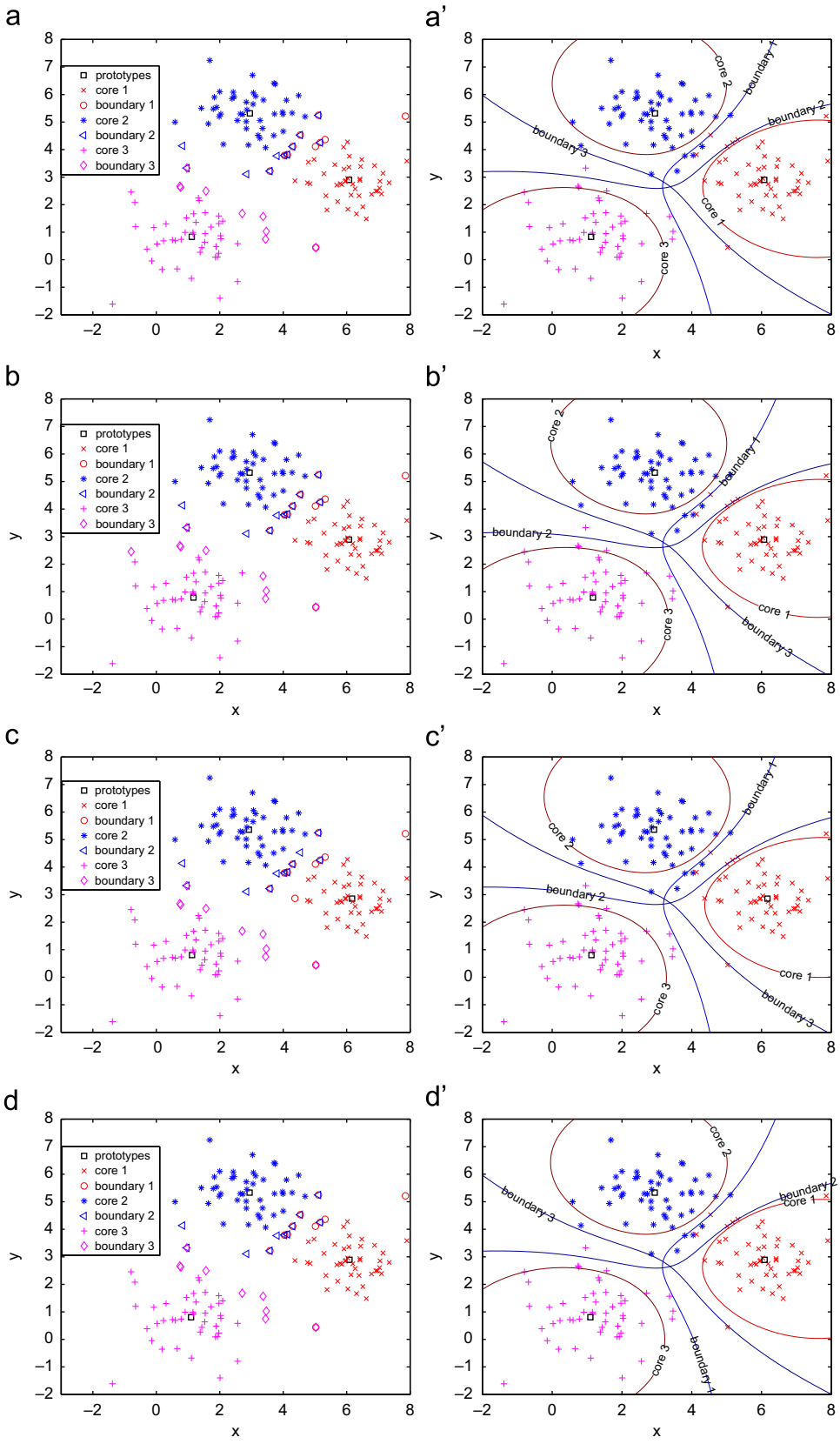

Fig. 5. Synthetic dataset I—Visualization of regions and boundaries generated by different methods: (a) SCM; (b) SRCM; (C) SRFCM I; and (d) SRFCM II. The left column presents the classification of each object and the formed prototypes. The right column plots the approximation regions of each cluster.

Compared with FCM, the contribution produced by the objects in the core, boundary and exclusion regions are enhanced, reduced and eliminated, respectively. Among introduced methods, the SRFCM I exhibits the best performance as documented in Table 3. In addition, the computed time (in seconds) of FCM is less than shadowed set- and rough set-based methods. The

reason is that the optimization of α for each cluster in each iteration consumes extra time.

6.2. Synthetic dataset II

Synthetic dataset II comes with the two clusters of quite distinct cardinalities. The prototypes produced by each method are collected in Table 4 and visualized in Figs. 6 and 7, respectively.

The underlying characteristics of these data affect the validity of the clustering methods. The prototypes produced by FCM, SCM, SRCM and SRFCM II somewhat deviate from the anticipated positions of the representatives. Here the FCM algorithm performs quite poorly. Although SCM, SRCM and SRFCM II show some improvement, the results are still not appealing, see Fig. 7(a, b and d). The results generated by SRFCM I are the best in terms of the location of the prototypes, see Fig. 7(c).

As shown in Fig. 7(a, b and d), the approximation regions obtained by SCM, SRCM and SRFCM II are not desirable. Some data objects are apparently partitioned into wrong areas, namely, three data objects that should belong to the cluster 1 are definitely assigned to the cluster 2. Inaccurate lower bounds and boundary regions will directly result in unsuitable prototypes. The best approximation regions of each cluster can be captured by SRFCM I, which can be observed in Fig. 7(c). Only a single object here has not been properly assigned to the clusters.

Different threshold values result in distinct approximation regions and these regions affect the prototypes and associated membership values. Following the principle of uncertainty balance in shadowed sets, the optimal threshold value of each cluster can be acquired. The comparative results are presented in Table 5 and the corresponding approximation regions of each cluster are shown in Fig. 7(right column). It can be observed that the core level of one cluster is the exclusion level of the other cluster. It reflects the duality property between approximation regions of the target concept and its complement in rough set methodology.

The validity indices of each method are compared in Table 6. SRFCM I exhibits a far better performance than other methods with respect to the available and newly proposed indices. Moreover, shadowed set- and rough set-based clustering methods, namely SRCM, SRFCM I and II, perform better than the generic SCM and FCM. It implies that the partition of approximation regions can reveal the nature of data structure and only the lower bound and boundary region of each cluster have *positive* contribution in the process of updating the prototypes. Though the execution time of shadowed set-based methods is longer than the one for the FCM method, they can also be realized in short time, as shown in Table 6.

6.3. UCI datasets

Eight UCI datasets are included in the experiments, namely Iris, Wine, Balance, Ionosphere, Wisconsin, Bupa liver, Vehicle and Heart data. The results of comparative analysis are shown in Tables 7–10. From the experimental results, the following conclusions can be drawn:

(1) The shadowed set-based *C*-means clustering methods perform far better than the FCM itself. The improvement can be attributed to the fact that the objects are divided into different regions (segments), which helps capture better the overall topology of the data.

(2) The shadowed set- and rough set-based clustering methods (namely SRCM, SRFCM I, and SRFCM II) exhibit better performance than the generic SCM. Through the weighted approaches, the contribution of each approximation region to the formation of the prototypes and the clusters can be properly quantified.

(3) It can be found that even though the computing time required to run FCM is less than the one required by shadowed set-based methods, FCM cannot provide sound results for all

Table 2
Comparative analysis for selected threshold values—synthetic dataset I.

	α_1	α_2	α_3
SCM	0.33848	0.30733	0.30318
SRCM	0.34086	0.30741	0.31123
SRFCM I	0.33568	0.31872	0.30445
SRFCM II	0.3398	0.30851	0.30228

Table 3
Validity indices—synthetic dataset I.

	PBM	XB	DB	Q	Time
FCM	13.935	0.087073	0.54924	0.32015	0.016
SCM	14.584	0.08383	0.53793	0.31256	0.265
SRCM	14.336	0.083301	0.53584	0.31229	0.141
SRFCM I	**14.816**	**0.07819**	**0.52427**	**0.30824**	0.25
SRFCM II	14.726	0.08263	0.53409	0.31049	0.156

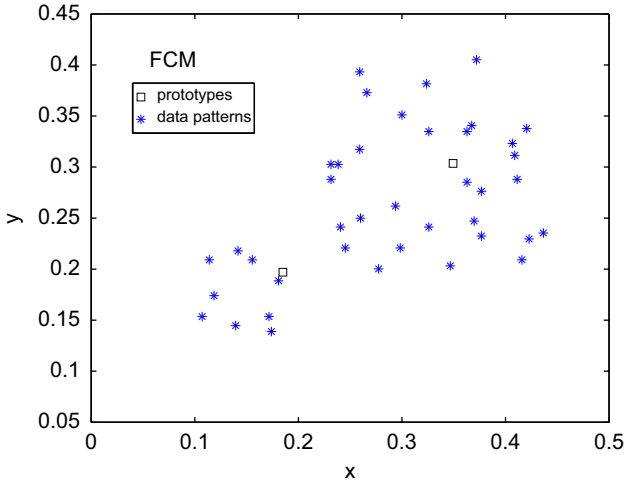

Fig. 6. Prototypes formed by the FCM—synthetic dataset II.

Table 4
Prototypes obtained for the synthetic dataset II.

	FCM		SCM		SRCM		SRFCM I		SRFCM II	
Prototype 1	0.34945	0.30355	0.34406	0.30148	0.34417	0.30156	0.34538	0.30014	0.34419	0.30161
Prototype 2	0.18505	0.19699	0.17922	0.19097	0.17782	0.19026	0.15579	0.18094	0.17779	0.1902

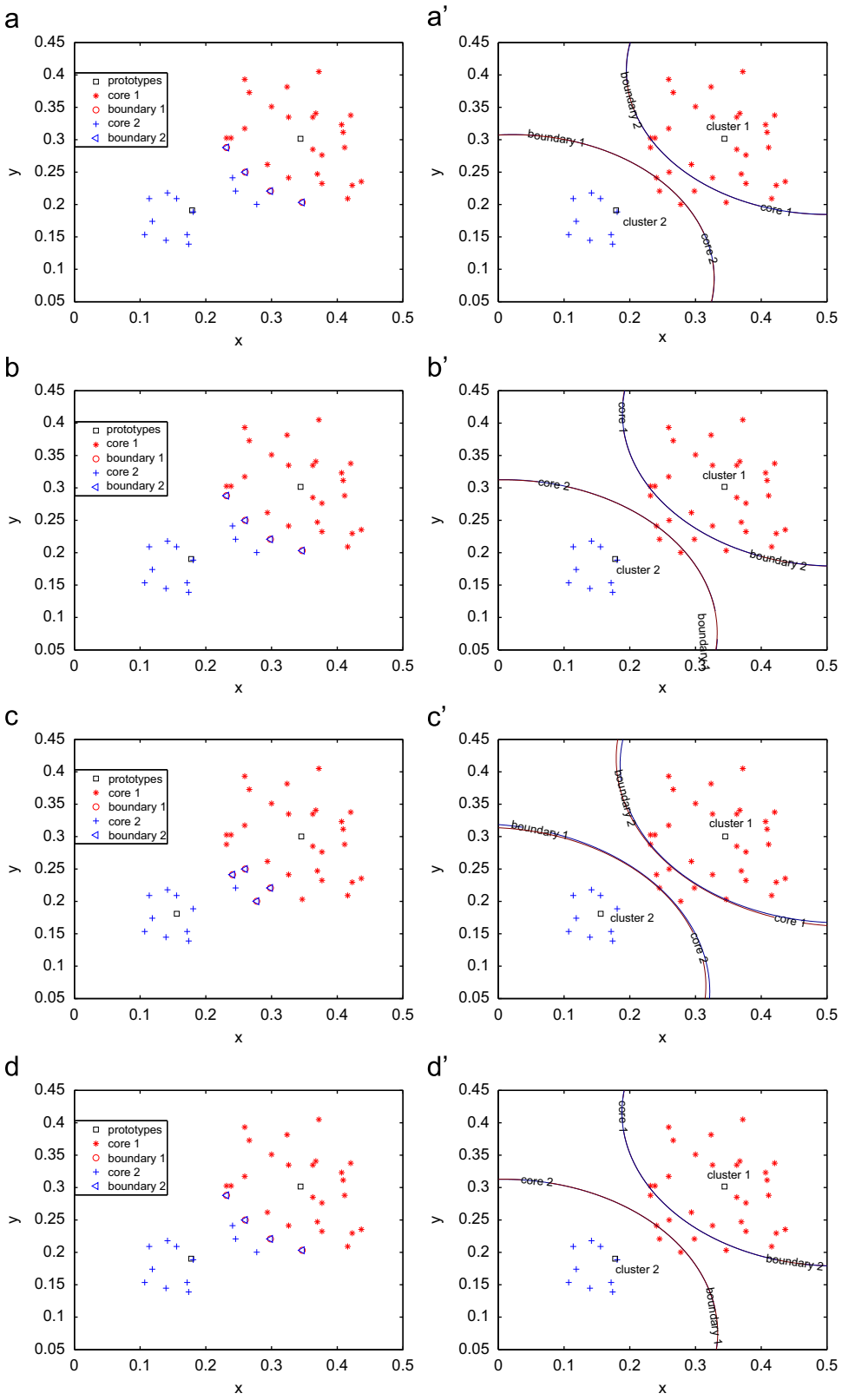

Fig. 7. Synthetic dataset II—approximation regions produced by different clustering methods: (a) SCM; (b) SRCM; (C) SRFCM I; and (d) SRFCM II. The left column presents the classification of each object and the formed prototypes. The right column plots the approximation regions of each cluster.

real-world data. However, the shadowed set-based methods can also be executed in short time along with better performance. Especially, they exhibit the capability on the Balance and Heart data which cannot be effectively handled by FCM.

(4) The SRFCM I method exhibits the best performance when being compared with the results produced by other methods and the quality is assessed by the validity indices (with exception of the PBM index reported for the Iris data). Here the advantages of

Table 5
Comparative results of thresholds for synthetic dataset II.

	α_1	α_2
SCM	0.32827	0.32785
SRCM	0.33629	0.33559
SRFCM I	0.33438	0.32549
SRFCM II	0.33629	0.33655

Table 6
Values of the validity indices—synthetic dataset II.

	PBM	XB	DB	Q	Time
FCM	4.7038	0.1295	0.74606	0.7441	0.015
SCM	4.7996	0.12655	0.69913	0.74393	0.078
SRCM	4.8829	0.12441	0.69171	0.7412	0.078
SRFCM I	**5.9473**	**0.10075**	**0.61455**	**0.72311**	0.14
SRFCM II	4.8886	0.12427	0.6913	0.74091	0.078

fuzzy sets, rough sets and shadowed sets are integrated in the SRFCM I. The membership grades make the proposed notion applicable to deal with overlapping partitions, as the concept of approximate regions can handle the uncertainty and vagueness arising from the boundary regions, and the optimization process in the shadowed sets make the method robust to outliers, so that the approximation regions of each cluster can be determined accurately and the obtained prototypes approach to the desired locations. Although the SRFCM II has the same properties, the experimental results demonstrate that the objects within the lower bound of a cluster should have different influence on this cluster and the calculations of the corresponding prototype when the shadowed sets are incorporated to the method.

7. Conclusions

The value of the threshold that determines the approximation regions in rough set-based clustering methods is crucial in the

Table 7
Validity indices for Iris and Wine data.

	Iris ($C=3$)					Wine ($C=3$)				
	PBM	XB	DB	Q	Time	PBM	XB	DB	Q	Time
FCM	40.722	0.18017	0.75115	0.89867	0.015	2.3341	6.6822	2.6731	7.6861	0.016
SCM	43.081	0.14207	0.68622	0.85073	0.125	4.099	1.4808	1.2496	6.873	0.172
SRCM	**43.946**	0.12821	0.66264	0.84523	0.172	4.28	1.2074	1.1454	6.8344	0.094
SRFCM I	42.933	**0.12614**	**0.65972**	**0.8388**	0.25	**4.4756**	**1.1419**	**1.1034**	**6.774**	0.266
SRFCM II	43.913	0.12868	0.66342	0.84424	0.203	4.3144	1.2257	1.1481	6.8178	0.125

Table 8
Validity indices for Balance and Ionosphere data.

	Balance ($C=3$)					Ionosphere ($C=2$)				
	PBM	XB	DB	Q	Time	PBM	XB	DB	Q	Time
FCM	0.004805	145.14	37.657	3.9906	0.078	0.52844	0.83245	2.0598	27.626	0.047
SCM	1.054	0.29652	1.4742	2.6284	1.532	0.92234	0.47282	1.5328	26.201	2.64
SRCM	1.2211	0.25702	1.359	2.5637	1.578	1.0149	0.42709	1.4544	25.992	0.203
SRFCM I	**1.226**	**0.24278**	**1.3083**	**2.5275**	4.281	**1.0963**	**0.39255**	**1.3946**	**25.883**	0.328
SRFCM II	1.1921	0.24792	1.343	2.5671	1.578	1.0226	0.42366	1.4481	25.973	0.203

Table 9
Validity indices for Breast cancer and Bupa liver disorders data.

	Breast cancer—Wisconsin ($C=2$)					Bupa liver disorders ($C=2$)				
	PBM	XB	DB	Q	Time	PBM	XB	DB	Q	Time
FCM	5.8505	0.11314	0.7599	3.7047	0.031	0.38034	0.82882	2.0251	4.9856	0.015
SCM	6.1424	0.10766	0.73838	3.6638	0.219	1.9217	0.16503	1.0422	4.1725	0.438
SRCM	6.1233	0.10788	0.7444	3.6441	0.344	2.3601	0.13128	0.94987	4.0806	0.344
SRFCM I	**6.4839**	**0.1016**	**0.7216**	**3.6069**	0.61	**2.4859**	**0.12484**	**0.92599**	**4.055**	0.453
SRFCM II	6.163	0.1072	0.74158	3.6423	0.234	2.3848	0.13016	0.94724	4.0824	0.313

Table 10
Validity indices for Vehicle and Heart-Statlog data.

	Vehicle ($C=4$)					Heart-Statlog ($C=2$)				
	PBM	XB	DB	Q	Time	PBM	XB	DB	Q	Time
FCM	11.447	2.0111	1.7879	8.4717	0.359	3.20E−08	1.16E+07	8280.9	13	0.062
SCM	13.168	2.2233	1.59	7.2515	2.297	0.14162	2.5329	3.6541	11.748	1.266
SRCM	13.214	1.7829	1.4529	7.1238	1.578	0.07401	4.8255	5.067	11.631	1.5
SRFCMI	**13.469**	**1.2596**	**1.3387**	**7.103**	3.578	**0.20503**	**1.6686**	**2.9167**	**10.668**	0.297
SRFCMII	13.443	1.7359	1.4408	7.1122	2.532	0.065	5.5288	5.4449	11.852	1.531

determination of prototypes so that they are reflective of the structure of the data. By engaging the optimization supported by the shadowed set constructs, the threshold is automatically acquired in rough set-based clustering methods. As a result, from the perspective of each cluster, all objects to be clustered are divided into three components. Since the lack of knowledge regarding global relationships over all objects caused by the individual absolute distance in RCM or individual membership degree in RFCM is diminished, the comparative accurate lower bound and boundary region of each cluster can be captured. The effectiveness of the proposed notion is demonstrated by experimenting some synthetic as well as real-world datasets.

The complex characteristics of data distribution cannot be fully analyzed by only a single methodology. The performance of the approach can be improved by integrating the available methodologies since all of them have their own merits and share a strong nature of complementarities. To comprehensively reveal the capabilities of the proposed hybrid methods, some possible applications need further investigation along with their theoretical basis. In addition, the proposed notion is implemented on static data in this study. How to utilize them for analyzing time-varying data is a challenging task to study in the future.

Acknowledgements

The authors are grateful to the anonymous referees for their valuable comments and suggestions. This work was supported by the National Natural Science Foundation of China (Serial nos. 60475019, 61075056, 60970061) and The Research Fund for the Doctoral Program of Higher Education in China (Serial no. 20060247039).

Appendices

Tables 11 and 12.

Table 11
Synthetic dataset I.

Index	x	y	Index	x	y	Index	x	y
1	3.4654	1.0284	51	4.3005	4.1126	101	4.1331	3.8171
2	0.18183	−0.35795	52	3.2691	5.0615	102	4.7904	3.5472
3	1.9709	1.5886	53	2.8449	3.1098	103	7.0358	2.6073
4	2.7071	1.6787	54	3.0342	6.7051	104	6.0239	4.2794
5	1.9233	0.097911	55	3.9913	5.2905	105	4.9716	3.6432
6	0.50687	0.71848	56	1.6382	4.9401	106	5.9741	2.8693
7	2.007	−1.3978	57	3.9792	5.3606	107	6.8801	3.4199
8	0.026907	0.57245	58	2.1137	4.1912	108	6.4778	3.6395
9	1.5408	1.708	59	2.6438	5.2954	109	7.1449	3.1369
10	0.22675	0.68569	60	2.8572	5.4483	110	6.4397	3.3893
11	−0.6546	1.1999	61	1.6835	7.2433	111	5.847	2.7293
12	0.96241	3.3294	62	2.3897	5.8543	112	5.6877	2.7087
13	0.74402	2.6719	63	3.4468	4.068	113	6.1394	2.3481
14	0.65946	−0.33351	64	0.58813	5.0003	114	6.9228	2.4913
15	1.3807	0.27017	65	4.6895	5.1976	115	5.3792	1.8063
16	1.3257	2.2398	66	2.0317	5.291	116	6.0726	3.3622
17	0.30765	1.3045	67	4.4889	5.9636	117	7.3409	2.7984
18	1.1544	0.99009	68	1.9834	5.2015	118	6.4162	2.8808
19	−0.29511	0.38581	69	0.8242	4.1379	119	4.5391	4.5291
20	1.9716	1.3114	70	2.4263	6.0836	120	6.0867	3.7606
21	−0.79335	2.4571	71	2.4508	5.9118	121	5.1691	4.2533
22	−0.14213	−0.044124	72	4.0014	4.6632	122	6.7847	2.3801
23	−0.077785	1.166	73	3.7482	6.3785	123	6.9698	2.4914
24	0.76866	2.6298	74	3.7242	5.2954	124	5.9035	4.0919
25	1.1055	−0.6792	75	3.7019	4.5108	125	6.1917	2.748
26	1.4313	0.43745	76	3.8053	3.7716	126	5.9178	3.29
27	2.0267	0.86831	77	2.8349	5.7234	127	7.8527	5.2132
28	1.2491	1.3578	78	1.9978	6.1029	128	6.2697	2.5735
29	−0.67869	2.0814	79	3.5817	3.2178	129	4.7633	2.7942
30	0.59862	0.69194	80	5.1113	5.2466	130	6.39	2.3224
31	1.5265	0.63547	81	3.0601	6.0595	131	5.2843	3.1262
32	3.4427	0.74753	82	3.051	5.6477	132	5.7176	2.111
33	0.78195	0.73888	83	3.7101	6.411	133	6.037	1.8349
34	1.1765	1.667	84	2.9047	4.1669	134	7.0517	2.8526
35	0.9488	1.5228	85	1.7963	5.477	135	6.5098	1.9205
36	2.0917	1.4005	86	3.2633	4.4135	136	5.0275	0.44225
37	1.729	1.1857	87	2.8392	4.7623	137	6.6205	1.4836
38	1.1644	0.92502	88	2.8347	5.0178	138	7.0982	2.3903
39	0.91015	0.9903	89	4.1027	5.7959	139	5.3469	2.3245
40	1.9698	0.23419	90	3.3762	5.8006	140	7.8969	3.585
41	−1.383	−1.6127	91	1.4191	5.5039	141	4.844	2.7594
42	2.5592	−0.79511	92	1.9022	5.5345	142	6.0335	2.955
43	2.0344	0.76323	93	3.2103	4.5213	143	4.369	2.8642
44	1.5424	0.95165	94	2.5544	5.0587	144	6.3115	1.6647
45	1.5648	2.4951	95	2.7306	5.2782	145	5.3215	4.3646
46	1.884	0.47951	96	1.8279	5.9025	146	5.7982	2.7656
47	2.5657	0.58422	97	3.1277	5.9385	147	5.6266	3.4521
48	3.3639	1.569	98	4.0411	3.7908	148	5.9532	2.841
49	1.3545	2.149	99	4.2827	5.3268	149	6.4163	2.9329
50	1.8619	0.092487	100	3.349	4.7691	150	5.0174	4.1097

Table 12
Synthetic dataset II.

Index	x	y	Index	x	y	Index	x	y
1	0.11866	0.17398	16	0.23157	0.28801	31	0.40899	0.3114
2	0.11406	0.20906	17	0.25922	0.31725	32	0.3629	0.3348
3	0.15553	0.20906	18	0.32604	0.3348	33	0.26613	0.37281
4	0.18088	0.1886	19	0.32373	0.38158	34	0.23848	0.30263
5	0.17166	0.15351	20	0.36751	0.34064	35	0.24078	0.24123
6	0.1394	0.14474	21	0.40668	0.3231	36	0.4159	0.20906
7	0.10714	0.15351	22	0.3629	0.28509	37	0.41129	0.28801
8	0.14171	0.21784	23	0.36982	0.24708	38	0.3001	0.351
9	0.17396	0.13889	24	0.32604	0.24123	39	0.2775	0.2002
10	0.23157	0.30263	25	0.29378	0.2617	40	0.26	0.25
11	0.25922	0.39327	26	0.29839	0.22076			
12	0.37212	0.40497	27	0.34677	0.20322			
13	0.42051	0.33772	28	0.37673	0.23246			
14	0.43664	0.23538	29	0.42281	0.22953			
15	0.24539	0.22076	30	0.37673	0.27632			

References

[1] J. MacQueen, Some methods for classification and analysis of multivariate observations, in: L. Lecam, J. Neyman (Eds.), Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, 1967, pp. 281–297.
[2] J.C. Bezdek, Pattern Recognition With Fuzzy Objective Function Algorithms, Kluwer Academic Publishers, Norwell, MA, USA, 1981.
[3] Z. Pawlak, Rough sets, International Journal of Information and Computer Science 11 (1982) 314–356.
[4] P. Lingras, C. West, Interval set clustering of web users with rough k-means, Journal of Intelligent Information Systems 23 (1) (2004) 5–16.
[5] S. Mitra, H. Banka, W. Pedrycz, Rough-fuzzy collaborative clustering, IEEE Transactions on Systems, Man, and Cybernetics (Part B) 36 (2006) 795–805.
[6] L.A. Zadeh, Towards a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic, Fuzzy Sets and Systems 90 (1997) 111–117.
[7] W. Pedrycz, Granular computing—the emerging paradigm, Journal of Uncertain Systems 1 (2007) 38–61.
[8] P. Maji, S.K. Pal, Rough set based generalized fuzzy c-means algorithm and quantitative indices, IEEE Transactions on Systems, Man, and Cybernetics (Part B) 37 (2007) 1529–1540.
[9] S. Mitra, An evolutionary rough partitive clustering, Pattern Recognition Letters 25 (2004) 1439–1449.
[10] G. Peters, Some refinements of rough k-means clustering, Pattern Recognition 39 (2006) 1481–1491.
[11] G. Peters, M. Lampart, R. Weber, Evolutionary rough k-medoid clustering, in: Transactions on Rough Sets VIII, Lecture Notes in Computer Science, vol. 5084, 2008, pp. 289–306.
[12] W. Pedrycz, Shadowed sets: representing and processing fuzzy sets, IEEE Transactions on Systems, Man, and Cybernetics (Part B) 28 (1998) 103–109.
[13] S. Mitra, W. Pedrycz, B. Barman, Shadowed c-means: integrating fuzzy and rough clustering, Pattern Recognition 43 (2010) 1282–1291.
[14] M.K. Pakhira, S. Bandyopadhyay, U. Maulik, Validity index for crisp and fuzzy clusters, Pattern Recognition 37 (2004) 487–501.
[15] D.L. Dubes, D.W. Bouldin, A cluster separation measure, IEEE Transactions on Pattern Analysis and Machine Intelligence 1 (1979) 224–227.
[16] X.L. Xie, G.A. Beni, Validity measure for fuzzy clustering, IEEE Transactions on Pattern Analysis and Machine Intelligence 13 (1991) 841–847.
[17] Z. Pawlak, A. Skowron, Rudiments of rough sets, Information Sciences 177 (2007) 3–27.
[18] W. Pedrycz, Knowledge-Based Clustering: From Data To Information Granules, Wiley & Sons INC, Publication, 2005.
[19] W. Pedrycz, From fuzzy sets to shadowed sets: interpretation and computing, International Journal of Intelligent Systems 24 (2009) 48–61.
[20] T.W. Liao, A clustering procedure for exploratory mining of vector time series, Pattern Recognition 40 (2007) 2550–2560.
[21] J. Yu, Q.S. Cheng, H.K. Huang, Analysis of the weighting exponent in the FCM, IEEE Transactions on Systems, Man and Cybernetics (Part B) 34 (2004) 634–639.
[22] W. Pedrycz, A dynamic data granulation through adjustable fuzzy clustering, Pattern Recognition Letters 29 (2008) 2059–2066.
[23] A. Frank, A. Asuncion, UCI Machine Learning Repository [〈http://www.ics.uci.edu/ml〉], University of California, School of Information and Computer Science, Irvine, CA, 2010.

Jie Zhou received his M.E. degree in Computer Science and Technology from Central South University, Changsha, China, in 2007. He has been a Ph.D. candidate of the Department of Computer Science and Technology, Tongji University, Shanghai, China, since 2007. He is currently a visiting student in the Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Canada. His current research interests include rough set theory, data mining and soft computing.

Witold Pedrycz is a professor and Canada Research Chair in the Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Canada. He is also with the Systems Research Institute of the Polish Academy of Sciences. He is actively pursuing research in Computational Intelligence, fuzzy modeling, pattern recognition, knowledge discovery, neural networks, granular computing and software engineering. He has published vigorously in these areas. He is an author of eleven research monographs and numerous journal papers in highly reputable journals. Dr. Pedrycz has been a member of numerous program committees of international conferences in the area of Computational Intelligence, Granular Computing, fuzzy sets and neurocomputing. He currently serves as an Associate Editor of IEEE Transactions on Fuzzy Systems, IEEE Transactions on Neural Networks. He is also on editorial boards of over 10 international journals. Dr. Pedrycz is also an Editor-in-Chief of Information Sciences and IEEE Transactions on Systems, Man, and Cybernetics part A. He is the past president of IFSA and NAFIPS. He is a Fellow of the IEEE.

Duoqian Miao is a professor in the Department of Computer Science and Technology, Tongji University, Shanghai, China. He has published more than 50 papers in international proceedings and journals. His main research interests include rough set theory, pattern recognition, data mining and granular computing.