# ccp1

*by* 1 Ccp1

**Submission date:** 06-Feb-2018 01:55PM (UTC+0700)
**Submission ID:** 911860188
**File name:** ida_draft_version_djoko_010414.pdf (765.93K)
**Word count:** 13735
**Character count:** 68801

# Optimization overlap clustering based on the hybrid rough discernibility concept and rough K-Means

Article in Intelligent Data Analysis · July 2015

DOI: 10.3233/IDA-150746

CITATIONS
2

READS
35

**3 authors:**

Djoko Budiyanto Setyohadi
Universitas Atma Jaya Yogyakarta
**34** PUBLICATIONS **10** CITATIONS
SEE PROFILE

Azuraliza Abu Bakar
National University of Malaysia
**170** PUBLICATIONS **502** CITATIONS
SEE PROFILE

Zulaiha Ali Othman
National University of Malaysia
**105** PUBLICATIONS **432** CITATIONS
SEE PROFILE

Please cite as : D. B. Setyohadi, A. Abu Bakar, Z. A. Othman, Optimization overlap clustering based on the hybrid rough discernibility concept a k-means, Intelligent Data Analysis 19 (4) (2015) 795–823. 10.3233/IDA-150746
The final manuscript can be downloaded from :
http://content.iospress.com/articles/intelligent-data-analysis/ida746

# Optimization Overlap Clustering based onthe Hybrid Rough DiscernibilityConceptand Rough K-Means

Djoko Budiyanto Setyohadi, Azuraliza Abu Bakar, and Zulaiha Ali Othman
Data Mining and Optimization Research Group, Center for Artificial Intelligence Technology
Faculty of Information Science and Technology, University Kebangsaan Malaysia
Bangi, Selangor Darul Ehsan, 43000 MALAYSIA
djokobdy@gmail.com, {aab,zao}@ftsm.ukm.my

## Abstract

Technically, the problem of overlap in a dataset is viewed as an uncertainty problem and is solved using a fuzzy set theoretical approach, specifically, fuzzy clustering. This approach is powerful but has some problems associated with it, of which the design of the membership function is the most serious. There are many different techniques for optimizing fuzzy clustering, including those based on similarity decomposition and centroids of clusters. Furthermore, the problem of overlap clustering is still being studied to improve its performance, especially with respect to the membership optimization. Rough set theory (RST) is the complement of fuzzy set theory and evidence theory, which use different techniques to address the uncertainty problem in overlap clustering. Considering the simplicity of the membership computation in RST, we propose an overlap clustering algorithm, which involves the use of the discernibility concept of RST to improve the overlap clusters as an existing variant of the overlap clustering algorithm. The experiment described here demonstrates that this new method improves the performance and increases the accuracy of clustering while avoiding the time complexity problem. The experiment uses five UCI machine learning datasets. The complexity of the data is measured using the volume of the overlap region and feature efficiency. The experimental results show that the proposed method significantly outperforms the other two methods in terms of the Dunn index, the sum of the squared errors and the silhouette index.

Keywords: overlap clustering, discernibility, RK-means, uncertain, rough membership

## 1.Introduction

Clustering is a data mining (machine learning) technique used to assign data elements to related groups without advance knowledge of the group definitions. The goal of clustering is to assign similar objects to the same clusters and dissimilar objects to different clusters. Dissimilarities are assessed on the basis of the attribute values that describe the objects; therefore, the characteristics of the data influence the clustering process. Distance measures are utilized in most clustering methods.

Because clustering is an unsupervised learning technique, the original data are not labeled by classes. The goal of most clustering methods is to partition an unlabeled dataset, e.g., $\{x_1, x_2, x_3, ...x_n\}$,with each object $x_i \in \Re_n$, where $\Re$ is a real number clustered into $C$ subgroups so that objects in the same cluster are characterized by the highest levels of similarity. Real-world data distributions often involve ambiguous or overlapping structures, which require a clustering method that allows the objects to be members of two or more clusters [2]. Previous studies have proposed a solution to the problem involving viewing an ambiguous object located on an overlap cluster as an uncertain object. An ambiguous object is one that can be a member of more than one cluster. The initial solution is implemented by using Fuzzy C-Means (FCM)[16]. FCM uses membership to represent the probability that an object belongs to each cluster, providing the flexibility to represent the probability of a data point belonging to more than one cluster at the same time.

FCM is a well-known fuzzy clustering algorithm. Many fuzzy clustering algorithms have been developed and used in various applications that involve uncertainty caused by overlap data processing, such as remote sensing and medical image processing [1]. However, FCM has two major drawbacks that diminish its performance. First, FCM is sensitive to outlier values of the clusters, and second, FCM can be easily trapped at local minima for both of these

drawbacks reduce the algorithm performance to produce a good partition cluster. In addition, centroid, prototype FCM, make inadequate algorithm to deal with non spherical cluster. Therefore, several extensions of the FCM algorithm have been proposed to improve its performance. Considering the purpose of clustering, the main goal of FCM is achieved by a fuzzy function that is used for cluster formation or partition. Furthermore, the partition is reflected by a membership value. Indeed, in the FCM, the objective of fuzzy clustering is to find the appropriate partition based on the membership value. This objective can be achieved by developing a membership function or optimization algorithm [31].

Several extensions of FCM have been developed, although the basic computation in FCM performed by Euclidean distance is only suitable for clusters that are spherical in shape whereas many non spherical datasets, which is geometrical overlap, are provided and required a good classification algorithm [36]. It is naturally that the real world data is the cluster formation not only uncertain, incomplete but also can have various different shapes, such as ellipsoids, lines, and quadratics and so on. Two important issues associated with FCM are examined in this paper. The first issue is the optimization of the membership value, which affects the performance of cluster partitioning especially by using initial seed to lead the algorithm produce the good cluster [3][4]. The second issue is the development of an alternative approach of membership computation to address various overlap shapes in the cluster partitions in the dataset. The various overlap shapes are caused by geometrical complexity dataset. The second issue is important since it will increase the level difficulties of classification, and reduce the performance of classification [21].

We propose a hybrid clustering algorithm to address these two issues, i.e., optimization and various shapes of overlap clusters which are caused by characteristics of geometrical complexity dataset. The more overlap dataset, the more ambiguous of objects assignment be. In ambiguous data processing problems, a clustering algorithm must be able to address various shapes of the cluster partition. RST is a soft-computing method that has been proven capable of addressing the problem of ambiguous data processing [28]. The performance of RST depends on the approximation addresses, which are normally associated with a set of attributes, depending on the granularity of knowledge, and are determined by the indiscernibility relation. Problems arise because RST cannot be used directly to develop clustering algorithms due to the complexity problem [32] however RST has been reported successfully to solve problem in categorical clustering [14]. RST has the potential to be used for clustering as a result of certain advantages its features offer. RST can be used to extend the clustering algorithm such as Rough K-Means (RKM)[29]. RKM is an extension of K-Means that uses RST to solve vague data processing problems.

In contrast to FCM, RKM is performed by separating an overlap object from the crisp cluster only. Indeed, RKM is specific to interval clustering, and RKM is not addressed to solve the overlap clustering; thus, in RKM, the membership concept should be extended within the boundary region to add a fuzzy concept similar to FCM. In this paper, we propose a new hybrid overlap clustering algorithm, referred to as Rough K-Means Discernibility (RKMD). This algorithm addresses both optimization and various types of cluster formation issues by developing rough membership computation to produce better partition in overlap dataset. First, RKM is employed and optimized using the initial seed concept to generate the appropriate interval cluster as the main outcome [4]. Second, using the discernibility computation, we calculate rough membership degree as the probability of the uncertain object belonging to each cluster. In addition, both the initial seed and the membership calculation are based on the discernibility computation.

The rest of this paper is organized as follows. Section 2 reviews related research on the overlap clustering problem, focusing on variations of fuzzy clustering algorithms and hybrid soft-computing techniques for clustering. Section 3 introduces important concepts of the RST and Rough K-Means clustering algorithms. The proposed method, which consists of two main phases, is discussed in Section 4. Experiments and concluding remarks are given in Sections 5 and 6, respectively.

## 2. Related Research

Overlap measurement is a component of data complexity analysis, which concerns the study of the degree to which patterns can be extracted from a dataset, and the performance of a classification algorithm depends on the characteristics of the dataset [21][36]. In this context, recent research on data complexity analysis was reviewed to characterize the relationship between the complexity of a dataset and the performance of a classifier algorithm [18][21]. Based on this relationship, in our study, we refer to the overlap and uncertainty as the complexity of the dataset. Although there are many ways to measure the complexity of a dataset, these measures are limited to the geometric or topological properties of the class distributions, such as the consideration of nonspherical shapes and uncertainty in the dataset, e.g., measures of the volume of the overlap region (F2) and feature efficiency (F3).

The volume of the overlap region (F2) of the cluster is calculated based on the maximum and minimum values of each class feature. The computation can be defined using each feature ($f_i$) and the maximum $\max(f_i, c_j)$ and the minimum $\min(f_i, c_j)$ values for each class ($c_j$), as shown in Eq.1:

$$F_2 = \prod_i \frac{\max(\max(f_i, c_1), \max(f_i, c_2)) - \max(\min(f_i, c_1), \min(f_i, c_2))}{\min(\max(f_i, c_1), \max(f_i, c_2)) - \min(\min(f_i, c_1), \min(f_i, c_2))} \quad (1)$$

Overlap is found in almost every type of information, its multi-faceted form being the result of many different factors, including empirical measurement error, inadequate computational methods, and cognitive ambiguity. Overlap can be present in data at any point, potentially leading to serious inaccuracy within data processing, as this situation results in uncertain membership object within a cluster, and the cluster is not spherical in shape. In pattern recognition, dealing with overlap is a common problem because real data are frequently ambiguous. As a result, boundaries between many classes can be poorly delineated (see Figure 1).

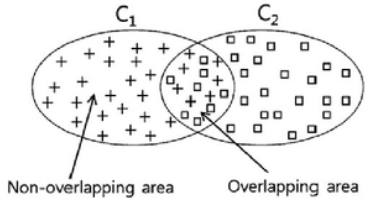
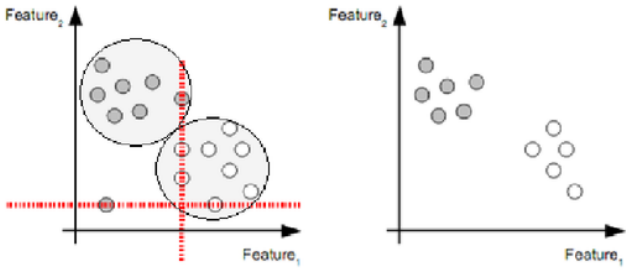



Figure 1.Overlap area as an uncertain/overlap dataset of clusters$c_1$ and $c_2$

From a different perspective, overlap feature efficiency (F3) focuses on feature/attribute overlap measurement (see Figure 2.a). This situation emerges when the cluster is nonspherical. Cluster overlap can form as shown in Figure 1, when clusters are not fully separated, or, as shown in Figure 2.a, when clusters are fully separated. The value of *F3* is calculated using the fraction of all remaining points of classes that are separable by that feature. As a result, the efficiency of each feature is the ratio of the remaining non-overlapping points to the total number of points. Suppose $p$ is all points of the same class; the largest feature efficiency of all features is taken as *F3*, as shown in Eq.2.

$$F_3 = \sum_p separable(p)$$
$$where\ separable(p) = \begin{cases} 1 & \text{if } p \text{ is separable by the feature} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Because the complexity of the dataset can influence the performance of the overlap clustering algorithm, both complexity measurements will be used to simulate the proposed method. Appropriate real-life datasets and a validation index are used to validate the performance of the overlap clustering. The details of the real-life dataset and the validation index used are described in Section 5 (experimental setup).




(a) (b)

Figure 2. The comparison of datasets and datasets with (a) without a feature overlap area (b)

## 2.1 Previous Development

Significant work has been performed to develop overlap clustering. In this section, we first provide a brief overview of the overlap problem and then follow recent developments in the improvement of overlap clustering based on FST, including improved membership computation, combination with other soft-computing approaches, and FCM optimization. The traditional FCM partitions a set of object data into a number of $c$ clusters based on the minimization of a quadratic objective function. The objective function to be minimized is as follows:

$$J_{FCM} = \sum_{i=1}^{c}\sum_{k=1}^{n}(\mu_{ik})^m \|x_k - v_i\|^2 \quad (3)$$

where $x_k$ : input data ($k = 1,\ldots, n$),

$v_i$ :the centroid of cluster $i$($i=1\ldots, c$), and

$\mu_{ik}$ :the fuzzy membership vector $x_k$ belonging to cluster $i$

$m$ ($m> 1$): the weighting exponent of fuzzy membership

The formula above is subject to the following constraint:

$$\sum_{i=1}^{c}\mu_{ik} = 1$$

Using the FCM approach, the overlap problem is solved because overlap clustering allows the condition in which an object belongs to two or more clusters. This approach is the main advantage of FCM, compared to hard clustering. FCM improves partition performance and reveals the classification data more reasonably. However, FCM has the well-known disadvantage of slow convergence [15]. In addition, the cluster formation characteristics may diminish the performance of the FCM algorithm. Several scholars report that most algorithms fail to clearly distinguish separated clusters, and therefore, their performance is often unpredictable when the degree of overlap of datasets is increased [18][21].

In FCM, the partition of the cluster is represented by the degree of membership. The following section describes several improvements in partitioning capability that have been reported in the fuzzy clustering literature as a form of overlap clustering. We review the development of FCM clustering in three subsections: extension of the function, FCM optimization, and hybrid optimization.

The appropriate partitioning, indicated by the best membership, is the objective of clustering. Many scholars have developed membership functions to address the FCM problem. Several extensions have been developed based on this method to address the problem of the data characteristics, which directly influence the performance of clustering. For example, the classical approach to the outlier problem in fuzzy clustering is to change probabilistic clustering to possibilistic clustering [27]. This approach is effective; however, it increases the sensitivity of the initial center, which leads to coincident clustering.

Another extension that has been proposed to address the nonspherical cluster problem is kernel fuzzy c-means [17][25]. Other improvements include the use of membership constraints [20] and changing the distance metric used for measurement [5]. A nonspherical dataset can also affect the density within a cluster. Therefore, the cluster might have different sizes and densities related to the cluster object. This situation leads to the cluster center problem, for which several algorithms have been proposed [6][24]. The RST has been used to extend clustering algorithm and has been reported in due to noisy, uncertainty [30][33] and categorical problem [14].

In FCM, the objective optimization approach is used to avoid the local optima problem, which is a common problem in the extension of c-means algorithms. The local minima problem can be avoided using the initial centroid method [3]. This method is implemented by feeding the initial centroid. Because this process involves complex computations, the use of an evolutionary algorithm (EA) is a common approach to optimizing fuzzy clustering. An EA is used to maintain the simplicity of a fuzzy clustering algorithm. The use of an EA has two advantages. First, an EA is able to learn the structure of the dataset and improve the performance of fuzzy clustering. Second, an EA is able to avoid the local minima problem in fuzzy clustering optimization. Indeed, an EA can be used to address complex optimization problems [31].

In addition to EA optimization, Genetic Algorithm (GA) and Particle Swarm Optimization (PSO) have been developed as popular FCM optimization methods [31]. In most cases, an initial population of randomly generated candidate solutions constitutes the first generation. A fuzzy membership fitness function is applied to the candidate solutions and any subsequent off spring until the optimal solution is achieved. Based on this characteristic, many optimizations of fuzzy clustering using PSO and GA have been developed [10][23][9][13].In comparing the performance of EA and PSO, several scholars have shown that PSO is superior, especially in terms of the speed of convergence [31].

A hybrid approach to addressing multi-clustering problems is promising. A hybrid approach combines the merits of multiple algorithms. Therefore, hybrid clustering may perform better than a single clustering approach [31][30]. A hybrid approach may be developed as a sequential algorithm, with one algorithm being applied after another algorithm is completed [26][13] or embedded in the objective function used[25].

Several hybrid fuzzy clustering approaches based on PSO have been developed. These approaches can increase the probability of finding the global optimum [31][35]. An example is Hybrid PSOKHM [13], a hybrid data clustering algorithm based on PSO and KHM that minimizes the sum over all data points of the harmonic average of the distance from each data point to all the centers using PSO and enables an increase in speed to avoid the local optima problem.

Recently, a hybrid fuzzy clustering particle swarm optimization approach, namely, FCM–FPSO [13] has been developed. This approach seeks to improve the convergence of FCM and the performance of FPSO while avoiding the local optima problem. FCM–FPSO algorithm can be divided in two stages; FPSO is aimed to find the best initial centroid than followed by FCM which is aimed to optimize overlap clustering using Eq.3. The objective of this algorithm is that the process is fast converge and produce better partition. We can conclude that the developments of fuzzy clustering rely on the fuzzy set theory to develop membership degree computation which is used to produce the best partition of dataset. Therefore, the new model computation of overlap clustering algorithm is needed.

2.2 FPSO

Local optima problem is the common problem c-means clustering due to the use random initial seed. Solving this problem, PSO is one of a population-based optimization. PSO has been used for many optimization problems. In c-means PSO can be applied in order to avoid local optima problem. Let says $X$ is position particle it can represented as matrix $n \times c$ of membership value $\mu_{nc}$ as below:

$$X = \begin{vmatrix} x_{11} & \dots & x_{1c} \\ \dots & \dots & \dots \\ x_{n1} & \dots & x_{nc} \end{vmatrix}$$

Iteration is performed to improve the partition by updating both positions and velocities as matrix operation below:

$$V(t+1) = w \times V(t) + (c_1 r_1) \times (par\_best((t)) - X(t) + (c_2 r_2) \times (glob\_best((t)) - X(t) \quad (4)$$

$$X(t+1) = X(t) \oplus V(t+1) \quad (5)$$

Due to the $\mu_{nc} = [0,1]$ the result of updated membership is normalized. Moreover the updated result is evaluated by objective function $f(x) = \frac{K}{J_m}$ . The process is repeated until termination value is achieved.

Let the dataset has $n$ objects $o = \{o_1, o_2, \dots o_n\}$ and $j$ cluster $c = \{c_1, c_2, \dots c_j\}$, P = particle of PSO, , $c_1$ and $c_2$ acceleration constants of PSO

The algorithm of FPSO can be writen as below

1. Initialize $P$, $w$, $c_1$ and $c_2$
2. Create swarm with $P$ particles
3. Initialize $X$,$V$, *par_best* for each particle and *glob_best* for the swarm
4. Calculate cluster center using for each particle Eq. 6

$$z_j = \frac{\sum_{i=1}^{n} \mu_{ij}^{m} o_j}{\sum_{i=1}^{n} \mu_{ij}^{m}} \quad (6)$$

5. Calculate fitness for each particle using $f(x) = \frac{K}{J_m}$
6. Calculate *par_best* for each particle and *glob_best* the swarm
7. Update velocity matrix each particle Eq.4
8. Update position matrix each particle Eq.5
9. If terminating condition is not achieved go to step 4

### 2.3 Hybrid FCM-PSO

Hybrid FCM-PSO [13] is an overlap clustering algorithm which is sequentially performed by FCM and FPSO clustering. The FPSO produce the partition and its centroid, and these results are fed for FCM. The use of centroid which is resulted by FPSO, is quite similar with the use of initial seed approach in c-means clustering for optimization [3][4]. The detail of the algorithm is described as below:

FPSO

1. Initialize $P$, $w$, $c_1$ and $c_2$
2. Create swarm with $P$ particles
3. Initialize $X$,$V$, *par_best* for each particle and *glob_best* for the swarm
4. Calculate cluster center using for each particle Eq. 6
5. Calculate fitness for each particle using $f(x) = \frac{K}{J_m}$
6. Calculate *par_best* for each particle and *glob_best* the swarm
7. Update velocity matrix each particle Eq.4
8. Update position matrix each particle Eq.5.
9. If terminating condition is not achieved go to step 4 otherwise do FCM phase

FCM

10. Calculate cluster centre for each particle using Eq. 6
11. Calculate Euclidian distance for each particle using Eq. 5
12. Update membership function $\mu_{ij}$ using Eq. 7

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c} \left( \frac{d_{ij}}{d_{ik}} \right)^{\frac{2}{m-1}}} \quad (7)$$

13. Calculate *pbest* for each particle and *gbest* the swarm
14. Repeat FCM if terminating condition is not achieved otherwise do next
15. Repeat from the beginning if Hybrid FCM-PSO terminating condition is not achieved

According to the algorithm, initial seed approach is used to avoid local optima problem in FCM (second phase). The applied concept is that initial seed should be closer with the final centroid. FPSO is performed to get the best

centroid by optimizing FCM. The expectation of this method is that the FCM in second phase will be able to get best partition and fast convergence.

## 3.Preliminaries

In this section, we review previous research i.e. RKM and discernibility RST which are used to develop the new overlap clustering. RKM is an approach to separating the overlap dataset from the crisp dataset, and discernibility is a concept associated with RST used to address the uncertainty problem in overlap clustering.

### 3.1 Rough K-Means Clustering

The most important issue addressed in rough set theory (RST) is the idea of imprecise knowledge. In this approach, knowledge is considered imprecise if it contains imprecise concepts. Imprecise concepts can be defined approximately by employing two precise concepts: lower and upper approximations [28]. Using these concepts, [29] Lingras proposed the Rough K-Means (RKM) algorithm, which addresses the problem of vague data. RKM's capability to cluster vague data comes from the integration of rough set theory with K-Means clustering. Whereas in the original K-Means approach, the cluster is viewed as a crisp cluster, in RKM, the cluster is viewed as an interval cluster. The object is divided into the lower approximation, where the object is certainly a member of the cluster, and the boundary region, where the object is a member of more than one cluster. In addition, if the dataset has an outlier, it will occur in the boundary region. RKM is able to address the outlier problem [11].

In contrast to FCM clustering, RKM yields a distinct result. FCM assigns membership to define whether an object belongs within a cluster, whereas RKM focuses on distinguishing between crisp and vague objects. A crisp object $x$ is assigned to the lower approximation of the cluster $\underline{appr}(c_l)$, whereas a vague object is assigned to the boundary region $bnd(c_l)$. The boundary region is the difference between the lower approximation and the upper approximation of the cluster resulting from the RKM clustering algorithm [29]. In this approach, the process of RKM clustering involves computing centroids based on the centroid of the lower approximation (the crisp area) and the boundary region. The membership of each area is determined using a relative distance measure, which is useful for reducing the influence of outliers [11].

Many researchers have successfully used the RKM algorithm to address vague data in various areas [22][7]. The good performance of RKM is due to the capabilities of RST, especially when the algorithm separates the crisp data from the vague data. Despite its advantages, RKM has two drawbacks: (1) a problem with numerical stability, in which the RKM equation requires that each cluster must have at least one member, and (2) the local optima problem, which is caused by the random initial seed used as the initial centroid determining the clustering result. Several researchers have solved these problems by improving the original RKM [4][8][11]. Peters (2006)[11] suggests the use of the ratio of distances instead of the differences between distance similarities. The differences vary based on the values in input vectors. Miao (2007) [8] uses angle measurements to determine the members of clusters and avoid empty clusters. With respect to the local optima problem, in all of the extensions of the K-Means algorithm, the problem can be avoided and the performance of the rough clustering algorithm can be enhanced by initial seed computation [4] .

However, considering the algorithm's ability to separate vague data from crisp data, RKM is seen as a powerful algorithm for clustering vague data, although it should be extended when RKM is used to address overlap clustering, as with FCM. Based on the advantages of RST in uncertain data processing, we enhance the capability of RKM using discernibility computation. This extension addresses the optimization problem and the inability of RKM to perform overlap clustering by adding membership values to the objects. The rest of this section describes discernibility, which will be used to develop the hybrid algorithm.

### 3.2 Discernibility of Rough Set Theory

The foundation of RST is that every object in the universe can be associated with some information. The associations are performed by the indiscernibility concept which is defined relative to a specified set of attributes. Furthermore, elementary set, composed by all indiscernible objects, forms the granule of knowledge about the universe. The unification elementary set is referred to as crisp set and otherwise is vague set as shown in Figure 3 below

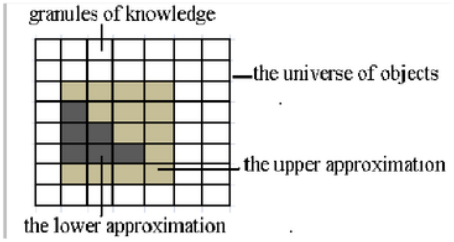



Figure.3. A Rough Set

Granularity comes from the rectangular grid and both crisp and vague set. It can be related by indiscernibility concept therefore the indiscernibility relation is one of the important properties of RST. Suppose that $IS=(U,A,V,f)$ is an information system, where $U=\{U_1,U_2,...,U_{|U|}\}$ is a finite non-empty set and universe object space, $A=\{a_1,a_2,...,a_{|A|}\}$ is the finite non-empty set of attributes. $V=\bigcup_{a\in A} V_a$, $V_a$ is the domain of the attribute $a$, and $f:U\times A\to V$ is an information function for $\forall a\in A,\forall x\in U\ \ f(x,a)\in\forall a$, which are points of the attribute value of each object in $U$. Each subset $B\subseteq A$ of attributes determines the indiscernibility $IND(B)$, which can be defined as shown in Eq.8 below:

$$IND(B)=\{(x,y)\in U\times U:\forall a\in B,\ f(x,a)=f(y,a)\} \tag{8}$$

Indiscernibility forms the equivalence class set. The equivalence relations induce a partitioning of the universe, meaning that all equivalence classes are disjointed and that their union is equal to the full universe of the set, and a partition also induces an equivalence relation. The strong indiscernibility relation with respect to $B$ is denoted by $IND(B)$. The two objects in $U$ satisfy $IND(B)$ if, and only if, they have the same values for all attributes in $B$. The quantitative discernibility relation $DIS(B)(x_i,x_j)$ is defined as the complement of a quantitative indiscernibility, as shown in Eq.9.

$$DIS(B)(x_i,x_j)=1-IND(B)(x_i,x_j) \tag{9}$$

To perform the above function, the decision/information table is discretized to construct the discernibility matrix table.

3.3. Rough Membership

Both fuzzy and rough set theory has been developed to deal with vagueness problem. Fuzzy set theory uses fuzzy membership which represents of gradualness of knowledge whereas rough membership in RST uses granularity of knowledge which is performed by the indiscernibility relation. The relation among elements of the universe $U$ and its equivalence class can be represented as a rough membership value $(\mu_x)$. The rough membership is calculated by using a relative quantifies a membership objects into a given set.

The value of $(\mu_x)$ reflects the approximation of the uncertainty property of the elements in a set. Formally, the rough membership is more general, as it reflects subjective knowledge about elements of the universe from fuzzy membership whereas the fuzzy membership indicates the conditional probability of the object belonging to the set [28]. Depending on the purpose of the clustering, rough membership is comparable to the membership degree value in fuzzy clustering, as $(\mu_x)$ can be interpreted as pertaining to the degree to which the element belongs to cluster $U$. Let $x\in X\subseteq U$ for any elementary granule $I(x)$ in subset $X$, and let $|.|$ be the cardinality. The probability that object $x$ belongs to cluster $X$ can be expressed as shown in Eq.10.

$$\mu_x^X:U\to[0,1] \tag{10}$$
$$where:\ \mu_x^X=\frac{|I(x)\cap X|}{|I(x)|}$$

Eq.10 shows that $\mu_x^X$ of the objects rely on the discernibility between objects in the cluster. Thus $\mu_x^X$ is not equal with fuzzy membership. However it can be used to represent the degree of belonging objects. In a discernibility-based data analysis, the traditional method of handling an attribute with a value set that is totally ordered is to partition its original value set into intervals and treat this discretized attribute as a categorical variable using the appropriate discretization approach. Furthermore, the rough membership can be used to compute the probability and to decide the overlap objects into proper class [28][34][12].

In the computation of discernibility, the discernibility matrix stores the sets of attributes that discern the values of the pairs of objects. Based on its characteristics, discernibility in rough set theory is seen as being capable of handling overlap clustering even when RST is used for unsupervised classification. The next section presents the proposed discernibility theory for overlap clustering, which is divided into two phases, namely, initial seed optimization, RKM Clustering and the incorporation of uncertainty in the vagueness area.

## 4.Proposed Method

Currently, overlap clustering is only viewed as an uncertainty problem. In this section, we propose an overlap clustering algorithm that incorporates the uncertainty property into the vagueness properties. Vagueness properties are applied using RKM to separate the crisp from the vague objects. The uncertainty properties of RST are used in the membership computation process when the algorithm addresses the vague objects in the boundary region. Because the original RKM uses a random initial seed to determine the initial number of clusters, it tends to become stuck at local optima. We overcome this problem using initial seed optimization [4]. We employ initial seed optimization based on RST to develop the discernibility classification using the uncertainty properties of RST. Figure 4 presents the design of a complete overlap clustering of the proposed method.

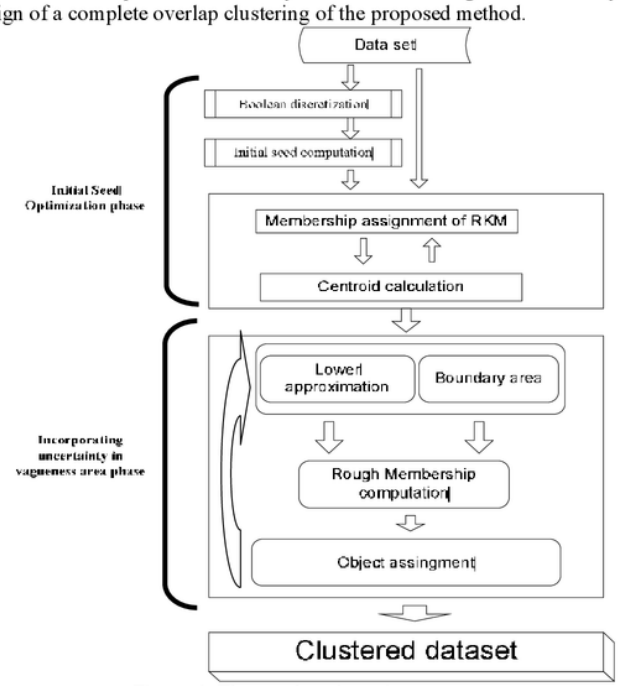



Figure 4. The overlap clustering process

### 4.1. Initial Seed Computation

K-Means and its extensions do not guarantee the optimization of clustering because the random selection of initial clusters leads to the local optima problem. The initial seed plays an important role in avoiding the local optima problem with this algorithm. In our proposed method, this concept is applied using the concept of the area of the initial seed. Correct determination of the area of the initial seed should avoid the local optima problem. To determine the area of the initial seed or initial centroid, Boolean discretization is used to partition the area into two parts in each dimension. The initial centroids are determined, and both centroids must have a sufficient distance, as selected from the points that have a high degree of discernibility. Figure 5 illustrates an example how the initial seed

is performed in for two dimensions and two clusters dataset. The area of the first and second initial centroids is used to determine the selected area of the initial seed.

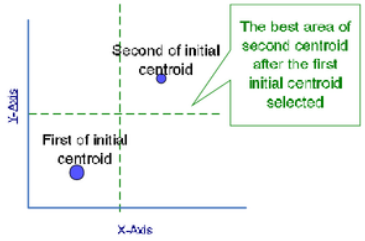



Figure 5. Area of the initial seed of a two-dimensional dataset

Suppose that $X = \{x_1, x_2, x_3, ..., x_n\}$ are objects in a dataset and that $\alpha_A(x_i, x_j)$ is the discernibility degree of objects $x_i, x_j$. The initial seed $S_i$ is the set of objects within the area dataset which has appropriate discernibility degree. This condition reflects that objects are suitable for the initial centroid in the RKM algorithm. This area has a high discernibility degree among the pairs of objects within the dataset. Figure 5shows an example of initial seeds $S_i$ and $S_i$ represented by pairs of objects in the first initial seed area and the second initial seed area. For$k$ clusters within dataset $X$, the $S_i$ can be calculated as shown below:

$$S_i = \left\{ (x_{i_1}, x_{i_2}, ..., x_{i_k}) \;\middle|\; \alpha_A(x_{i_1}, x_{i_2}, ..., x_{i_k}) \geq \theta \cdot max\left(\alpha_A(x_{i_1}, x_{i_2}, ..., x_{i_k})\right) and \;\; x_i \in X \right\} \tag{11}$$

$$where \;\; \alpha_A(x_{i_1}, x_{i_2}, ..., x_{i_k}) = \bigcup_{x_i \in X} \{min(\alpha_{i_m}, \alpha_{i_n})\}$$

$$and \;\; (\alpha_{i_m}, \alpha_{i_n}) = DIS(x_m, x_n) = \frac{|\{a \in A | I_a(x_m) \neq I_a(x_n)\}|}{|A|} \tag{11a}$$

$S_i$ is a set that can be computed using Eq. 11. The largest possible value of the discernibility of two objects is one (1), which means that the two objects are fully dissimilar, and the smallest possible value is zero (0), which means that the two objects are fully similar. However, the higher the density of the distribution is, the lower the discernibility values among objects will be due to the basic principle that objects that are used as initial seeds should be discernible from each other. We can choose the objects that have maximum discernibility. $\theta$ is a multiplication factor with a value between 0 and 1 that is used to control the area of the initial seed. $\theta = 0$ indicates the discernibility degree among objects which can be used as an initial seed that started from zero. Thus, this condition is equal with random initial seed since all of the objects can be used as an initial seed. For this purpose, $\theta$ can be chosen as 0.75 or more to exclude distant objects.

4.2. RKM clustering

The initial seed computation (Eq.11) produces the pair of objects that can be used as an initial centroid in the RKM algorithm [29]. The initial centroid is used to assign the membership of an object based on a *threshold* parameter that indicates whether the object belongs to the lower approximation of the cluster $\underline{appr}(c_l)$ or the boundary region $bnd(c_l)$. Let $x_l$ be an object in the dataset, and let $d(x_l, v_k)$ and $d(x_l, v_j)$ be the Euclidian distances between object $x_l$ and clusters $c_k$ and $c_j$, respectively. The membership of $x_l$ can be defined as follows:

i. If $\dfrac{\max(d(x_l, v_k), d(x_l, v_j))}{\min(d(x_l, v_k), d(x_l, v_j))} \leq threshold$ and $\min(d(x_l, v_k), d(x_l, v_j)) \neq 0$ then $x_l \in bnd(c_k)$ and $x_l \in bnd(c_j)$

ii. If $\dfrac{\max(d(x_l, v_k), d(x_l, v_j))}{\min(d(x_l, v_k), d(x_l, v_j))} > threshold$ and $\min(d(x_l, v_k), d(x_l, v_j)) = d(x_l, v_k)$ then $x_l \in \underline{appr}(c_k)$

iii. If $\frac{\max(d(x_l,v_k),d(x_l,v_j))}{\min(d(x_l,v_k),d(x_l,v_j))} > threshold$ and $\min(d(x_l,v_k),d(x_l,v_j)) = d(x_l,v_j)$ then $x_l \in \underline{appr}(c_j)$

iv. If $d(x_l,v_j) = 0$ then $x_l \in \underline{appr}(c_j)$

v. If $d(x_l,v_k) = 0$ then $x_l \in \underline{appr}(c_k)$

It should be noted that the approximation space $A$ is fully constructed based on the value assigned to the *threshold* parameter. $A$ is not defined on the basis of any predefined relation for the set of objects. The outcome of the previous rule is the lower approximation and the boundary region of each cluster within the dataset. Next, the centroid of each cluster is refined based on all of the objects of both areas.

$$v_l = \begin{cases} \omega_{low}\frac{\sum_{x\in\underline{appr}(c_l)} x_l}{|\underline{appr}(c_l)|} + \omega_{bnd}\frac{\sum_{x\in bnd(c_l)} x_l}{|bnd(c_l)|} & for\ \underline{appr}(c_l)\neq\varnothing \quad and \quad bnd(c_l)\neq\varnothing \\ \omega_{low}\frac{\sum_{x\in\underline{appr}(c_l)} x_l}{|\underline{appr}(c_l)|} & for\ \underline{appr}(c_l)\neq\varnothing \quad and \quad bnd(c_l)=\varnothing \\ \omega_{bnd}\frac{\sum_{x\in bnd(c_l)} x_l}{|bnd(c_l)|} & for\ \underline{appr}(c_l)=\varnothing \quad and \quad bnd(c_l)\neq\varnothing \end{cases} \tag{12}$$

The RKM algorithm described above depends on three parameters: $\omega_{low}$, $\omega_{bnd}$, and *threshold*. These parameters represent the rough treatment dataset and require some experimentation to obtain appropriate results from rough clustering. However, we can use two important characteristics to guide the selection of parameter values. The first characteristic is that the larger the threshold value is, the greater the probability is that outliers will be located in boundary region; however, a large threshold can result in a single cluster [11]. The second characteristic is represented by the weighting components $\omega_{low}$ and $\omega_{bnd}$ in Eq.12. Both of weighting components can be used to control the centroid relocation $v_i$.

The threshold value used in the membership assignment reflects how the RKM algorithm accommodates the overlap of objects as an overlap area, which is represented by a boundary region $bnd(c_l)$. If the ratio of the distances from objects $l$ to cluster $j$ and cluster $k$ is greater than the threshold value, then the object is considered similar to the closest cluster, whereas if the ratio is less than the threshold value, then the object is vague or an overlap object because it is considered similar to both of the two clusters. The threshold value will influence the width of the overlap space of the cluster. The result of this phase is the lower approximation of cluster $l$ $\underline{appr}(c_l)$ and its boundary region $bnd(c_l)$. The details of this computation are described in the complete example in the next section.

4.3 Incorporating Uncertainty in the Vagueness Area

The crisp set is a set that has been previously labeled by the RKM process. Let $S$ be a clustered information system in $U$, the universe set, and let $c$ be the number of clusters.

$$S = (U, A, V, f) \tag{13}$$

1. $U$ is a non-empty finite set of objects.
2. $A$ is a non-empty finite set of attributes. $A$ is limited to the crisp cluster and is further classified into two disjoint subsets of condition attributes $C$ and decision attributes $D$. In this algorithm, $D$ is generated with the lower approximation that results from the RKM process.
3. $V = \cup V_a$, where $a \in A$, $V_a$ is the domain of the attribute $a$.
4. $f : U \times A \rightarrow V$ is an information function that associates a unique value of each attribute with every object belonging to $U$.

In the RKM algorithm, $D$, the decision attribute in the information system $S$, is divided into two areas, namely, the crisp object area and the vague object area. The vague object area is located in the boundary region and has no decision attribute because the objects belong to more than one cluster. With respect to the result of

$x \in \underline{appr}(c_I)$ and $x \in bnd(c_I)$, given an information system, $S = (U, A)$, $X \subseteq U$, and $B \subseteq A$. The operation assigned to every $X \subseteq U$ with respect to $\underline{B}$ as the lower approximation of $\underline{X}$ is defined as follows:

$$\underline{B} = \bigcup_{x \in U} \{B(x) : B(x) \subseteq X\} \tag{14}$$

With respect to $\underline{B} \approx \underline{appr}(c_i)$, Eq. 14 reflects the approximation in RST and represents association, which is used as a foundation to perform rough membership computation. To perform rough membership computation, we must make two assumptions.

(1) The crisp objects, the objects in the lower approximation, are viewed as trained data. This assumption is valid because the data have been labeled by the RKM process.

(2) Each element in the boundary region has at least one indiscernibility attribute with respect to the objects in the lower approximation/crisp cluster. This assumption is also valid because of the use of the initial seed approach in RKM.

Because both assumptions are valid, the membership computation based on the discernibility relation can be performed. The discernibility concept required by the Boolean function is regarded as the foundation of granular computation, which requires the discernibility matrix as a representation of the atomic granules of a set. Conventionally, the discernibility matrix commonly requires Boolean computation characteristics only, whereas in our approach, the Boolean computation is required to accommodate the distance concept. Combining both requirements, we propose the new discernibility matrix described below. Let an information system have $n$ objects and $j$ attributes with $\{x_1, x_2, x_3, ...x_n,\} \subset A x \Re$. Each object is discretized based on data characteristics. The discernibility matrix can be generated using Eq.11.

$$d_{ikj} = \begin{cases} 1 & if \ x_{ij} \geq cut_{jk} \\ 0 & otherwise \end{cases} \tag{15}$$

*i:* the number of objects

j: the total number of attributes in information system

k: the maximum partition of the attributes resulting from the discretization method; each partition will have a value partitioned as $d_{ijk}$

Using the above formulation, given a crisp cluster $\underline{appr}(c_i)$ and an overlap cluster that contains object $x, \forall x \in bnd(c_i)$, the numerical characterization of the degree to which an object $x$ belongs in cluster $c_i$, relative to the knowledge represented by an attribute/feature set, is provided by the rough membership function $\mu_x^{c_i}$ (see Eq.10).After the discernibility matrix of all of the vectors in the universe is converted using Eq.15, $\forall x \in \underline{appr}(c_i)$ can be used as the foundation of the calculation $\mu_x^{c_i}$ of $x, \forall x \in bnd(c_i)$.

The classical membership fuzzy function and its extension use the point-to-point distance as a basis for the computation (see Eq.7).There are two main disadvantages to this approach. First, if the partition shape of the dataset is nonspherical, the resulting partition will be forced to have a spherical shape, which will lead to a misclassified cluster partition, although the algorithm is optimized using well-known global optimization approaches such as EA and PSO. When the cluster has more feature overlaps (see Figure 2.a), the use of conventional distance measures, such as the Euclidian distance, will not capture a nonspherical cluster caused by a feature overlap (F3). Second, the measurement similarity from point to point based on the distance within the overlap cluster will replace the appropriate centroid, especially in an overlap feature cluster [15][31][13]. Furthermore, the performance of the fuzzy clustering algorithm will be diminished. Thus, we propose the use of membership based on RST. The proposed membership function takes into account the relative similarity of each discerning object to the cluster. This function can be performed after the crisp cluster is found, and it can be completed easily using the discernibility calculation of the RKM process. Furthermore, discernibility is implemented to calculate the membership degree of the boundary object using the equivalence class.

4.4. Rough Membership computation

The membership of a vague object is related to the overlap measurement and the equivalence class. Furthermore, the equivalence class object $x(x^*)$ is represented by each discerning feature of the crisp cluster, which has a similar approximation in a set. Let the discretization performed be the approximation of feature $f$, and let $f_{d^*}$ denote the interval to which the feature of dataset $S$ belongs. This computation can be performed using the discernibility matrix:

$$\left[s^*\right] = \left\{(x, x') \in U^2 \middle| f_{s^*}(x) = f_{s^*}(x')\right\} \tag{16}$$

The membership of object $x$ corresponding to the discerning lower approximation of cluster $\underline{appr}(c_i)$ is denoted by $\mu_x^{c_i}$. Using the membership concept in RST, $\mu_x^{c_i}$ is computed as shown below:

$$\mu_x^{c_i}(x) = \frac{\left|\left[x_c^*\right] \cap c_i\right|}{\left|\left[x_c^*\right]\right|} \tag{17}$$

In rough set theory, $\mu_x^{c_i}$ reflects how discerning vector $x$ is related to crisp cluster $\underline{appr}(c_i)$. In the membership concept, this value is comparable to the conditional probability of $x$ belonging to cluster $\underline{appr}(c_i)$. Because this relation may violate the constraint of the uncertain membership concept, i.e., $\sum_{c=1}^{n} \mu_x^{c_i} = 1$, we should normalize $\mu_x^{c_i}$ as shown in Eq. 18.

$$\mu_x^{c_i} = \frac{\mu_x^{c_i}}{\sum_{c=1}^{n} \mu_x^{c_i}} \tag{18}$$

The membership value represents the probability of an object being a member of a cluster. Because of the influence of the crisp cluster,in the next iteration, the membership of the crisp cluster is changed. Thus, we assign object $x$ to the cluster $c_i$ that has the highest rough membership, as shown in Eq. 19 below.

$$if \ \ \mu_x^{c_i} = \max\left(\mu_x^{c_i}, i = 1,..c\right) \ then \ x \in c_i \tag{19}$$

Using the newer cluster, we repeat the process of rough membership computation until all of the objects in the boundary region have the same rough membership value.

Psedocode1 shows how the proposed method is performed.

Input : unlabeled overlap dataset
Output : clustered dataset

*Begin*
*Until all of the objects is discretize do*
*Assign 0 for objects below mean of attribute and otherwise*
*End Until*
*Calculate* $\alpha_A$
*Find Max_A=Max(*$\alpha_A$*)*
*Select objects to be member of* $S_i$ *using* $\alpha_A > 0.75 \times Max\_A$
*Select one of pair wise of set* $S_i$ *and use as an initial centroid*
*Until termination condition of RKM is achieved*
*Assign one of pair wise* $S_i$ *as an initial centroid*
*For i=1 to object_number do*
*Assign object into boundary area or otherwise*
*End for*
*Calculate a new centroid using Eq 12*
*Check termination condition*
*End Until*
*For i=1 to cluster_number of boundary_of region do*
*Calculate* $\mu_x^{c_i}$ of object
*Assign objects into the cluster which has the highest* $\mu_x^{c_i}$
*End For*
End

4.5. An Example

We present below an example of how the proposed method addresses the vague data based on the crisp cluster. In the first process, data is processed by using discernibility initial seed computation. This step is used to measure discernibility degree of objects dataset. Moreover, the proper objects, which are used as an initial seed, are selected based on discernibility degree.

4.5.1. Initial seed computation

Consider a dataset that contains two (2) clusters and two (2) attributes i.e., attribute 1 and attribute 2 (see Table 1.a). For initial seed purposes, Boolean discretization is performed to produce the discretization (Table 1.b).

Table1. Example dataset (a) and the Boolean discernibility table (b)

| (a) | | | (b) | | | |
|---|---|---|---|---|---|---|
| Object | Attribute 1 | Attribute 2 | Attribute 1 | | Attribute 2 | |
| 1 | 5.1 | 3.5 | 1 | 0 | 0 | 1 |
| 2 | 4.9 | 3 | 1 | 0 | 0 | 1 |
| 3 | 4.7 | 3.2 | 1 | 0 | 0 | 1 |
| 4 | 5.5 | 2.3 | 1 | 0 | 1 | 0 |
| 5 | 6.5 | 2.8 | 0 | 1 | 1 | 0 |
| 6 | 5.7 | 2.8 | 0 | 1 | 1 | 0 |
| 7 | 6.3 | 3.3 | 0 | 1 | 0 | 1 |
| 8 | 4.9 | 2.4 | 1 | 0 | 1 | 0 |
| 9 | 6.6 | 2.9 | 0 | 1 | 1 | 0 |
| 10 | 5.2 | 2.7 | 1 | 0 | 1 | 0 |
| 11 | 6.2 | 3.4 | 0 | 1 | 0 | 1 |
| 12 | 5.9 | 3 | 0 | 1 | 0 | 1 |

Eq. 11 and Eq. 11.a are used to calculate the discernibility degree between object $x_1, x_2$, and object $x_1, x_3$, as shown below:

$$\alpha_A(x_1, x_2) = \frac{|0+0+0+0|}{|4|} = 0$$

$$\alpha_A(x_1, x_3) = \frac{|0+0+0+0|}{|4|} = 0$$

Based on the example dataset, the discernibility degree among objects can be calculated. Thus the discernibility table can generated as shown in Table 2.

Table2. Discernibility table based on Boolean discretization

| $\alpha_A(x_i, x_j)$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | | | | | | | | | | | |
| 2 | 0 | 0 | | | | | | | | | | |
| 3 | 0 | 0 | 0 | | | | | | | | | |
| 4 | 0.5 | 0.5 | 0.5 | 0 | | | | | | | | |
| 5 | 1 | 1 | 1 | 0.5 | 0 | | | | | | | |
| 6 | 1 | 1 | 1 | 0.5 | 0 | 0 | | | | | | |
| 7 | 0.5 | 0.5 | 0.5 | 1 | 0.5 | 0.5 | 0 | | | | | |
| 8 | 0.5 | 0.5 | 0.5 | 0 | 0.5 | 0.5 | 1 | 0 | | | | |
| 9 | 1 | 1 | 1 | 0.5 | 0 | 0 | 0.5 | 0.25 | 0 | | | |
| 10 | 0.5 | 0.5 | 0.5 | 0 | 0.5 | 0.5 | 1 | 0.5 | 0.5 | 0 | | |
| 11 | 0.5 | 0.5 | 0.5 | 1 | 0.5 | 0.5 | 0 | 0.5 | 0.5 | 1 | 0 | |
| 12 | 0.5 | 0.5 | 0.5 | 1 | 0.5 | 0.5 | 0 | 0.5 | 0.5 | 1 | 0 | 0 |

There are many pairs of objects that have high discernibility values (greater than 0.75), i.e., 1. This means that we can use one of the pairs of objects as in initial seed to replace the initial random seed used in the RKM algorithm. The limitation on the initial seed space is intended to avoid zero-member problems and optimize the RKM clustering algorithm.

$$S_{initial\ seed} = \{(1,5),(1,6),..(7,8),..,(12,10)\}$$

4.5.2. RKM Process

The result of the first process is the proper objects which fit into criteria of initial seed. More over these objects are used as the initial seed of RKM. We believe that the proper initial seed able to lead RKM to produce good cluster. In this step this example will show how RKM able to separate vague object from boundary region by using proper initial seed. For example, object 7(6.3, 3.3) and object 8 (4.9, 2.4) are selected as an initial seed. Thus, the initial centroids in the RKM algorithm are $v_1$= object 7 and $v_2$= object 8. Suppose that the values of the RKM parameters are *threshold* =1.6, $\omega_{low} = 0.6$, and $\omega_{bnd} = 0.4$ .The following Euclidian distances between object $x_1$ and each centroid are calculated:

$$d(x_1,c_1) = 1.2166,\ d(x_1,c_2) = 1.1180$$

To calculate the distance ratio, the shortest distance is selected as the divisor, i.e., $d(x_1,c_2) = 1.1180$ :

$$\frac{1.2166}{1.1180} = 1.0881 \le 1.6$$

Because the ratio of the object is less than the threshold value, $x_1$ is assigned to the boundary region of cluster $c_1$. The 1st object is a vague object, based on the membership rule, and is located in the boundary region. This means that the similarity of the 1st object to cluster $c_1$ is comparable to the similarity of the 1st object to cluster $c_2$.

For the next object, i.e., $x_2$,

$$d(x_2,c_1) = 1.4318, d(x_2,c_2) = 0.6000$$

To calculate the distance ratio, the shortest distance is selected as the divisor, i.e., $d(x_2,c_2) = 0.6000$ , and the membership rule is applied to locate object $x_2$:

$$\frac{1.4318}{0.6000} = 2.3863 \ge 1.8$$

Unlike the 1st object, the similarity of the 2nd object to cluster $c_1$ is significantly different from the similarity of the 2nd object to cluster $c_2$. Because the 2nd object is closer to cluster $c_2$, this object is assigned to cluster $c_2$.

As a result, the lower approximation and the boundary region after the 1st iteration are as shown in Table 2.

Table3. The first result of iteration of the RKM algorithm

| | Lower bound of cluster | Boundary region of cluster |
|---|---|---|
| cluster (1): | $object_5$, $object_7$,$object_9$, $object_{11}$, $object_{12}$ | $object_1$, $object_4$,$object_6$ |
| cluster (2): | $object_2$, $object_3$,$object_8$, $object_{10}$ | $object_1$, $object_4$,$object_6$ |

The new centroid cluster $c_2$ is calculated based on Eq. 12. For the second cluster, the centroid of attribute 1 is as follows:

$$v_{attribut_1} = 0.6\frac{object_2 + object_3 + object_8 + object_{10}}{|4|} + 0.4\frac{object_1 + object_4 + object_6}{|3|}$$

$$v_{attribut_1} = 0.6\frac{4.9 + 4.7 + 4.9 + 5.2}{|4|} + 0.4\frac{5.1 + 5.5 + 5.7}{|3|} = 2.955 + 2.1733 = 5.1283$$

Using similar calculations, the centroid of attribute *2* of the second cluster is $v_{attribut_2} = 2.83$, and for the first cluster, the new centroid $v_{attribut_1} = 6.07$ and $v_{attribut_2} = 3.1$ .

Using the new centroids of cluster $c_1$ (6.07, 3.1) and cluster $c_2$ (5.1283, 2.83), the next iteration is performed using the assignment membership rule, followed by Eq.12. This iterative process continues until convergence is

achieved. We can also limit the number of iterations. After the final iteration, the results are as shown as Table 1. and Figure 6.

4.5.3. Assigning vague objects into crisp cluster

The RKM algorithm produces crisp cluster and its boundary region. Boundary regions represent the overlap area of dataset. In fuzzy clustering membership degree is used to make partition or to assign objects into cluster. Moreover, this process is aimed to calculate rough membership degree which is used to assign objects of boundary region into proper cluster. Using Eq. 15, we generate the discernibility matrix as the foundation of the membership computation.

Table4. The results of the RKM clustering algorithm

| object | attr1 | attr2 | Cluster |
|---|---|---|---|
| 1 | 5.1 | 3.5 | vague |
| 2 | 4.9 | 3 | 1 |
| 3 | 4.7 | 3.2 | 1 |
| 4 | 5.5 | 2.3 | vague |
| 5 | 6.5 | 2.8 | 2 |
| 6 | 5.7 | 2.8 | vague |
| 7 | 6.3 | 3.3 | 2 |
| 8 | 4.9 | 2.4 | 1 |
| 9 | 6.6 | 2.9 | 2 |
| 10 | 5.2 | 2.7 | 1 |
| 11 | 6.2 | 3.4 | 2 |
| 12 | 5.9 | 3 | 2 |

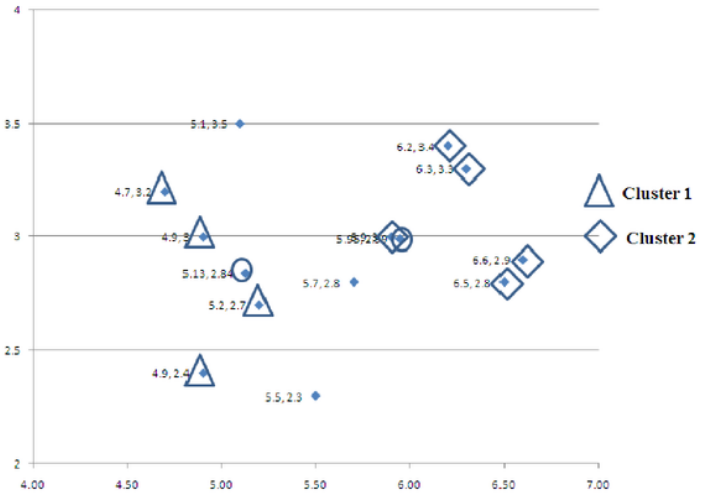



Figure 6. The results of the RKM clustering algorithm

Let the cut-offs of discernibility matrix conversion be as shown below.

Attribute 1: $cut_{11} = 0, cut_{12} = 5.1, cut_{13} = 5.7, cut_{14} = 6.1$

Attribute 2: $cut_{11} = 0, cut_{12} = 2.8, cut_{13} = 3.0, cut_{14} = 3.1, cut_{15} = 3.4$

Using Eq. 14, each value attribute of each object is converted

| | | | | | |
|---|---|---|---|---|---|
| 1st attributes of the 1st object 5.1 are converted to | 1 | 1 | 0 | 0 | 0 |
| 2nd attributes of the 1st object 3.5 are converted to | 1 | 1 | 1 | 1 | 1 |

After all the records are converted, we have the discernibility matrix shown below.

Table5.Discernibility matrix of the example dataset

| | $x_1^*$ | $x_2^*$ | $x_3^*$ | $x_4^*$ | $x_5^*$ | $x_6^*$ | $x_7^*$ | $x_8^*$ | $x_9^*$ | *cluster* |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | vague |
| 2 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 |
| 3 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 1 |
| 4 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | vague |
| 5 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 2 |
| 6 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | vague |
| 7 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 2 |
| 8 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| 9 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 2 |
| 10 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| 11 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 |
| 12 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 2 |

Using a lower approximation or crisp clusters $C^1$ and $C^2$, the membership of the 1st record is calculated as follows:

$$C^1 = \{x_2, x_3, x_8, x_{10}\}$$
$$C^2 = \{x_5, x_7, x_9, x_{11}, x_{12}\}$$
$$x_1^* = \{x_2, x_3, x_5, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}\}$$
$$x_2^* = \{x_5, x_7, x_9, x_{10}, x_{11}, x_{12}\}$$
$$x_3^* = \{x_5, x_7, x_9, x_{11}, x_{12}\}$$
$$x_4^* = \{x_5, x_7, x_9, x_{11}\}$$
$$x_5^* = \{x_2, x_3, x_5, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}\}$$
$$x_6^* = \{x_2, x_3, x_5, x_7, x_9, x_{11}, x_{12}\}$$
$$x_7^* = \{x_2, x_3, x_7, x_{11}, x_{12}\}$$
$$x_8^* = \{x_3, x_7, x_{11}\}$$
$$x_9^* = \{x_{11}\}$$

$$\mu_{C^1}^{x_1^*} = \frac{\left|[x_1^*] \cap [C^1]\right|}{|x_1^*|} = \frac{|x_2, x_3, x_8, x_{10},|}{|\{x_2, x_3, x_5, x_7, x_8, x_9, x_{10}, x_{11}, x_{12},\}|} = \frac{4}{9} = 0.4444$$

$$\mu_{C^2}^{x_1^*} = \frac{\left|[x_1^*] \cap [C^2]\right|}{|x_1^*|} = \frac{|x_5, x_7, x_9, x_{11}, x_{12}|}{|\{x_2, x_3, x_5, x_7, x_8, x_9, x_{10}, x_{11}, x_{12},\}|} = \frac{5}{9} = 0.5555$$

$$\mu_{C^1}^{x_2^*} = \frac{\left|[x_2^*] \cap [C^1]\right|}{|x_2^*|} = \frac{|x_{10}|}{|\{x_5, x_7, x_9, x_{10}, x_{11}, x_{12},\}|} = \frac{1}{5} = 0.2$$

$$\mu_{C^2}^{x_2^*} = \frac{\left|[x_2^*] \cap [C^2]\right|}{|x_2^*|} = \frac{|x_5, x_7, x_9, x_{11}, x_{12}|}{|\{x_5, x_7, x_9, x_{10}, x_{11}, x_{12},\}|} = \frac{4}{5} = 0.8$$

Similarly, the rough membership values corresponding to all $x^*$ can be computed. The final approximation can then be calculated as follows:

| | $f_1^*$ | $f_2^*$ | $f_3^*$ | $f_4^*$ | $f_5^*$ | $f_6^*$ | $f_7^*$ | $f_8^*$ | $f_9^*$ | $\sum^*$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $\mu_1^{c_1}$ | 0.4444 | 0.2 | 0 | 0 | 0.4444 | 0.2857 | 0.4 | 0.3333 | 0 | 2.1078 |
| $\mu_1^{c_2}$ | 0.5555 | 0.8 | 1 | 1 | 0.5555 | 0.7142 | 0.6 | 0.6666 | 1 | 6.8918 |

After normalization, the membership values of the record can then be computed using Eq.19, as shown below:

$$\mu_1^{c_1} = \frac{2.1078}{2.1078 + 6.8918} = 0.2342$$
$$\mu_1^{c_2} = \frac{6.8918}{2.1078 + 6.8918} = 0.7657$$

Because the membership value of the 2nd cluster is larger, the1st object is assigned to the 2nd cluster to perform the subsequent processing of vague objects.

There are two important advantages to using this algorithm in real data processing. First, the basic foundation of the proposed overlap clustering is hard k-means clustering. Thus, this method has less complexity and requires less computation time. Second, the use of the approximation concept of RST in the rough membership computation (see Eq.14) improves nonspherical overlap clustering performance (see Figure 1 and Figure 2.a).

## 5. Experiments

In this section, we assess the performance of the proposed method based on three index measures. In Section 5.1, the indexes and dataset used, based on the overlap characteristics, are described. Section 5.2 presents the results and discussion. We apply Eq. 15 using the entropy/MDL routine in the ROSSETA software[1]to obtain the cutoff of each partition of the dataset.

### 5.1. Experiment Design

In this study, the proposed method is tested using five publicly available datasets obtained from the UCI Machine Learning Data Repository. According to two issues of FCM, we select the UCI dataset which has been investigated on the level of difficulty based on geometrical complexity [36]. The volume of the overlap region (F2) and the feature efficiency (F3) are considered as geometrical complexity which represents level of overlap problem. Furthermore, the datasets are chosen based on their varying overlap and difficulty levels (low to high) in order to validate how overlap cluster influences the performance of overlap clustering algorithm. The datasets used in this study are the Haberman, Iris, Pima, Wine, and Wisconsin datasets.

Table 6 lists the geometrical complexity of each dataset used in this study that is taken from Ho (2002) [36]. The number of the attribute (#Attribute), the cluster (#Cluster), and the size of the dataset (#Data) are recorded. As discussed in Section 2, we use F2 and F3 to measure both the volume of the overlap region and the feature efficiency to determine the degree of overlap and the cluster complexity which represents the difficulty level of classification. The lower F2 is or the higher F3 is the more separable the data. The characteristics of the datasets used are shown in Table 3.

Table 6. Dataset complexity based on separability measures (F2 and F3)

| Dataset | #Attribute | #Cluster | #Data | F2 | F3 |
|---|---|---|---|---|---|

[1]The ROSETTA homepage [http://rosetta.sourceforge.net/] developed by Alexander Øhrn.

| | | | | | |
|---|---|---|---|---|---|
| Wine | 13 | 3 | 178 | 0.001 | 0.564 |
| Iris | 4 | 3 | 150 | 0.114 | 0.500 |
| Wisconsin | 9 | 2 | 683 | 0.217 | 0.350 |
| Pima | 8 | 2 | 768 | 0.251 | 0.651 |
| Haberman | 3 | 2 | 306 | 0.718 | 0.029 |

5.2. Validity indices

The performance of the clustering algorithm is commonly focused on how separate and how compact the result is. This validation is performed by evaluating the results of the clustering algorithm and using information that involves the vectors of the datasets themselves. In addition the visual inspection also performed to show the resulted cluster. Due to the dimension of the data, PCA is used to transform the data into two dimensions. The compactness and separation are measured using Dunn Index while homogeneity is validated using Sum Square Error. Silhouette is used to measure the performance of objects assignment. Rough membership is comparable with fuzzy membership but it is not equal [28]. Moreover these combinations are addressed on how the uses of rough membership of the proposed overlap clustering able to deal with overlap dataset. In addition, we add a visual inspection order to assess how well an algorithm provides users with a clear and intuitive understanding of the cluster and its structure [19]. Furthermore three internal validations and the purpose are described below:

i. The Dunn index ($D$) identifies the cluster sets that are compact and well separated. $U_{i,j}$ is the distance between cluster $i$ and cluster $j$ for $1 \le i, j \le c$ .

$$D = \min_j \left\{ \min_{i \neq j} \left\{ \frac{d(U_i, U_j)}{\max_k S(U_k)} \right\} \right\} \tag{20}$$

Using this index, the inter-cluster separation should be maximized, whereas the intra-cluster distances should be minimized. The larger the value of the Dunn index is, the better the clusters are separated. Dunn index is selected rather than Davies Bouldin Index since the inter-cluster separation of Dunn index relies on the minimum pair wise distance between objects in different clusters. This condition make Dunn index provides a rich and very general structure for different types of clusters.

ii. The homogeneity is considered, the second fitness evaluation is performed using the sum of the squared errors (SSE), as defined in Eq. 21. The homogeneity of the formed clusters is represented by the average Euclidean distance of the object to the centroid. The smaller the SSE is, the higher the quality of the clustering is.

$$\arg_{\min} \min \sum_{j=1}^{k} \sum_{x_j \in S_i} \left\| x_j - c_i \right\|^2 \tag{21}$$

iii. The last validation is performed using the Silhouette Index (SC), which measures the quality of the clusters in terms of the object assignments in the cluster. SC is calculated as shown Eq. 22, where $a(x)$ is the average distance from $x$ to all other objects in the similar cluster, and $b(x)$ is the average distance from $x$ to the objects in other clusters.

$$SC = \frac{1}{N} \sum_{i=1}^{N} s(x) \tag{22}$$

$$where \quad s(x) = \frac{b(x) - a(x)}{\max\{a(x), b(x)\}}$$

Based on the properties of $a(x)$ and $b(x)$, $s(x)$, the Silhouette Index (SC) represents how close an object to its cluster compared to how far it is from other clusters. Using this characteristic, we can evaluate whether $x$ is clustered properly by the overlap clustering algorithm. If the Silhouette Index is close to *+1*, $x$ is close to its cluster; if $s(x)$ is close to *-1*, $x$ is misclustered.

The performance of the proposed clustering algorithm, called RKMD, is compared with that of two fuzzy-based optimization algorithms, specifically, the FPSO algorithm and the Hybrid FCM-PSO hybrid optimization algorithm [13]. To validate the algorithms, $c$ the number of clusters is entered manually, the value is taken similar as the characteristics dataset. The hybrid optimization FPSO and Hybrid FCM-PSO is selected as comparator. Hybrid

FCM-PSO clustering algorithm is a two stage of overlap clustering. Integration PSO in FPSO of Hybrid FCM-PSO is used to produce initial centroid which is used as an initial centroid of FCM. Therefore it is comparable with the proposed algorithm which deploys indiscernibility computation for initial centroid. Shortly these experiments are purposed to measure the performance of overlap clustering especially on how initial seed and rough membership able to deal with various overlap dataset. The comparison of three algorithms is described as table 4 below.

Table 7. Comparison of FPSO, Hybrid FCM-PSO and RKMD

| No | | FPSO | Hybrid FCM-PSO | RKMD |
|---|---|---|---|---|
| 1 | Initial seed technique | Random | FPSO | Indiscernibility computation |
| 2 | Cluster prototype | Point (centroid) | Point (centroid) | Subset (crisp cluster) |
| 3 | Membership computation | Fuzzy Set | Fuzzy Set | Rough Set |

5.3 Results and Discussion

The proposed method, RKMD, is implemented using Java Netbeans 6.8 and a computer with a 1.85-Ghz Intel Core2Duo processor with3 GB of RAM and the Windows XP operating system. In our experiment, we use the previous result of or experiment [4], and we use *threshold*= 1.2, $\omega_{low} = 0.85$ and $\omega_{bnd} = 0.15$.The performance of RKMD is analyzed in terms of computation time and the results of a visual inspection and internal validation. In the internal validation, the proposed method is compared with two other algorithms, as explained in the previous section, namely, the FPSO and Hybrid FCM-PSO algorithms whereas these parameter required is set as Izakian (2011)[13] and both FPSO and Hybrid FCM-PSO run 100 times to produce independent experiments. The Dunn index, SSE, and silhouette index are used to measure the performance of the algorithms with respect to the five datasets.

5.3.1. Time computation

Time computation is the main problem when global optimization is applied in any clustering algorithm. The proposed algorithm relies on the use proper initial seed is addressed to reduce time computation. In addition the use RKM is useful to avoid time computation problem. The purpose of this experiment is addressed to validate the performance proposed algorithm with respect to the required time computation as listed in Table 4. below.

Table 8. Execution times of the RKMD, FPSO, and Hybrid FCM-PSO algorithms for the five datasets, in milliseconds (ms)

| No | Dataset | FPSO | Hybrid FCM-PSO | RKMD |
|---|---|---|---|---|
| 1 | Wine | 2652.50 | 2655.78 | 46 |
| 2 | Iris | 821.87 | 823.28 | 31 |
| 3 | Pima | 5063.26 | 5117.19 | 204 |
| 4 | Wisconsin | 4978.46 | 5023.28 | 157 |
| 5 | Haberman | 910.33 | 915.00 | 47 |

Table 4 shows the processing time of each algorithm for each dataset. The Hybrid FCM–PSO algorithm is developed by deploying the FCM algorithm in the FPSO algorithm. These characteristics can be viewed as the implementation of the initial seed concept in the FCM algorithm, which improves the performance of the entire c-means algorithm and its extension [3][4]. In this case, the initial seed is computed by FPSO and then fed into the FCM algorithm. The differences in the running times between the Hybrid FCM–PSO and FPSO algorithms are small (see Table8), whereas the performance is improved significantly (see Tables 9, 10, and 11). RKMD outperforms the other two algorithms, especially in terms of the execution time, without the clustering performance being diminished.

5.3.2. Visual Validation

Visualization is considered to be one of the most intuitive methods of cluster detection and validation and performs especially well in the representation of irregularly shaped clusters. Using visualization techniques allows us to evaluate, monitor, and guide the inputs, products, and processes of data mining. Therefore, the use of cluster visualization makes it possible to visualize the structure.

In many datasets, it is easy to become over helmed by the volume of measurements, which are represented by many features. Principal component analysis (PCA) can be used to reduce the dimensionality of the data so that a visual inspection can be performed more easily. Given a set of data, PCA finds the linear lower-dimensional

representation of the data such that the variance of the reconstructed data is preserved. Intuitively, PCA finds a low-dimensional hyperplane such that, when we project our data on to the hyperplane, the variance of our data is changed as little as possible. We visually inspect the RKMD and Hybrid FCM–PSO results as shown in Figures7.a–7.j below.

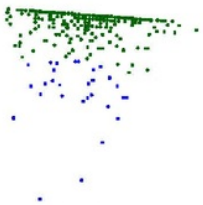


Figure 7.a. Haberman data RKMD

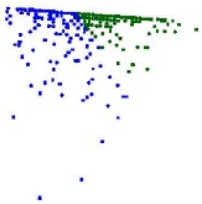


Figure 7.b. Haberman data Hybrid FCM–PSO

Figure 7.c. Iris data RKMD

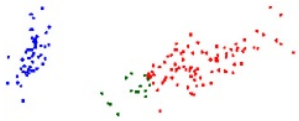


Figure 7.d. Iris data Hybrid FCM–PSO

Figure 7.e. Pima data RKMD

Figure 7.f. Pima data Hybrid FCM–PSO

Figure 7.g. Wine data RKMD

Figure 7.h. Wine data Hybrid FCM–PSO

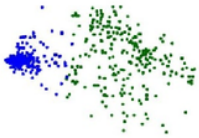


Figure 7.i. Wisconsin data RKMD

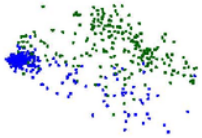


Figure 7.j. Wisconsin data Hybrid FCM–PSO

The clustered data of the five datasets are illustrated in Figures7.a–7j, in which the clusters are indicated by different colors. We used PCA, which preserves the data variance, to visualize the cluster distributions. The cluster distributions show that both of the algorithms are able to separate the overlap clearly, although they result in different clustering formations.

The Iris dataset is perhaps the clearest visualization of the overlap clustering problem. In the pattern recognition literature, Iris is a well-known dataset that is used to test classification algorithms. One of the clusters contains *Iris setosa*, and the other cluster contains both Iris Virginica and Iris Versicolor and is not separable without the species information Fisher used. Both algorithms perform well in separating the vague cluster (*Iris virginica* and *Iris versicolor*) and the crisp cluster (Iris Setosa), however the size of partition, which is produced by both of clustering algorithm, is different. However, in the visualization, the Hybrid FCM–PSO algorithm yields more unbalanced results (in the Iris dataset, each of the clusters include 50 data points).

5.3.3. Internal Validation

5.3.3.1. Compact and separateness

Table 5 presents a comparison of the Dunn index value for the proposed RKMD algorithm and the FPSO and Hybrid FCM–PSO algorithms. The results show that RKMD outperforms than the other two methods, as indicated by the significantly higher values for each dataset, which indicates that RKMD separates the clusters better than FPSO and Hybrid FCM–PSO. The clustering will be better if objects within cluster are similar whereas objects in different cluster should dissimilar. As the results in Table 9 show, the RKMD algorithm achieved the best Dunn Index values for all of the datasets. These results indicate that the clustering results of the RKMD algorithm are more compact and more separable than those of the other algorithms.

Table 9.Comparison of the Dunn index values

| | FPSO | | FCM–FPSO | | RKMD | |
|---|---|---|---|---|---|---|
| | average | Std dev. | average | Std dev. | average | Std dev. |
| Wine | 0.0359 | 0.0268 | 0.0313 | 0.0238 | 1.7345 | 0.0000 |
| Iris | 0.0605 | 0.0502 | 0.0521 | 0.0437 | 2.1542 | 0.0072 |
| Wisconsin | 0.0402 | 0.0205 | 0.0286 | 0.0146 | 1.5367 | 0.0000 |
| Pima | 0.0392 | 0.0243 | 0.0330 | 0.0208 | 2.0890 | 0.0008 |
| Haberman | 0.0580 | 0.0323 | 0.0558 | 0.0325 | 1.7796 | 0.0923 |

5.3.3.2. Homogeneity

The SSE presented in Table 10 show that the RMKD yields lower SSEs than FPSO and Hybrid FCM–PSO, which indicates that the clusters obtained are more compact. Initial seed is important in both c-means clustering. The homogeneity is one of important characteristics of the cluster. It represents the quality of clustering algorithm. Regarding the homogeneity problem, initial seed should be located in the region of each cluster. This condition is able clustering algorithm to produce cluster more homogen. SSE index can be used to measure this experiment and the result as below :

Table 10.Comparison of the SSE values

| | FPSO | | FCM–FPSO | | RKMD | |
|---|---|---|---|---|---|---|
| | average | Std dev. | average | Std dev. | average | Std dev. |
| Wine | 46099.94 | 306.44 | 43427.61 | 1589.39 | 16739.34 | 0.0000 |
| Iris | 289.73 | 1.68 | 275.10 | 8.93 | 102.03 | 0.2091 |
| Wisconsin | 5205.79 | 6.09 | 5145.17 | 52.27 | 3145.83 | 0.0000 |
| Pima | 74517.79 | 63.43 | 73470.86 | 804.57 | 54536.98 | 54.1013 |
| Haberman | 3557.94 | 4.88 | 3497.92 | 38.24 | 2734.11 | 197.1969 |

The homogeny cluster is important since the homogeny cluster means the contained more precise knowledge which useful when indiscernibility computation is performed. The lower SSE values of RKMD algorithm indicate that all of the objects within a cluster are closer to the centroid. This condition can be resulted also from location of the initial seed. The proper seeds should be relevant to the characteristics of clusters and it should be located at the homogeneity objects inner. If initial seed is not located in the similar partition with original centroid then the partitioning process will not able to produce partition which equal with original partition.

5.3.3.3. Objects Assignments

Our main goal is to improve overlap clustering, we measure the proper assignment of the objects using the silhouette index. The clustering is better when the value of this index is closer to +1. This value indicates that the object is assigned properly based on similarity of the attributes. Based on the index and complexity measurements (F2 and F3) obtained, the proposed method is shown to be able to address the overlap clustering problem by assigning the object to the correct cluster even when the complexity is increased, as shown in Figure 8 and Figure 9. This experiment show that the use of rough membership outperforms compared with fuzzy membership when it deals with overlap data. The performance comes from the use boundary region and discernibility computation. The aim of boundary region is reduce the overlap data processing problem in two ways. First, set is divided into region crisp and vague region. The valid crisp region is easier resulted by RKM. In addition the use of initial seed improves the compactness of crisp cluster [4]. Secondly, the indiscernibility is enable RKMD assign vague objects in boundary region into crisp cluster precisely. This performance comes from the provided knowledge which is contained in crisp cluster is valid. Thus the similarities of the objects in overlap clustering are is better since they are located into appropriate cluster accurately. This phenomenon is represented by silhouette index in Table 11.

Table 11.Comparison of the Silhouette Index values

| | FPSO | | FCM–FPSO | | RKMD | |
|---|---|---|---|---|---|---|
| | average | Std dev. | average | Std dev. | average | Std dev. |
| Wine | -0.0333 | 0.0120 | 0.4296 | 0.0121 | 0.5632 | 0.0000 |
| Iris | -0.0276 | 0.0091 | 0.4671 | 0.0688 | 0.5465 | 0.0000 |
| Wisconsin | 0.0030 | 0.0025 | 0.4383 | 0.1335 | 0.5192 | 0.0000 |
| Pima | 0.0025 | 0.0014 | 0.4084 | 0.0900 | 0.5099 | 0.0002 |
| Haberman | 0.0061 | 0.0030 | 0.3508 | 0.0649 | 0.4089 | 0.0135 |

An interesting result of the experiment is also shown in Figure 8 and Figure 9. Both figures represent the influence of the overlap condition on the performance of the algorithm, especially on the capability to assign objects. This condition can be interpreted as the capability of the proposed clustering to limit the influence that the overlap of the dataset has on performance. In other words, the proposed method is more stable or robust than the other methods. An anomaly is evident for the Pima dataset, as shown in Figure 9. This anomaly might be due the complexity

computation for the Pima dataset, which is different in terms of F2 and F3, as shown in Table 3.Based on the silhouette index values, it can be concluded that the proposed method is able to address the overlap clustering problem and assign objects to the correct clusters. A silhouette index value closer to 1 indicates that the objects are clustered correctly, as the members of the cluster are close to each other.

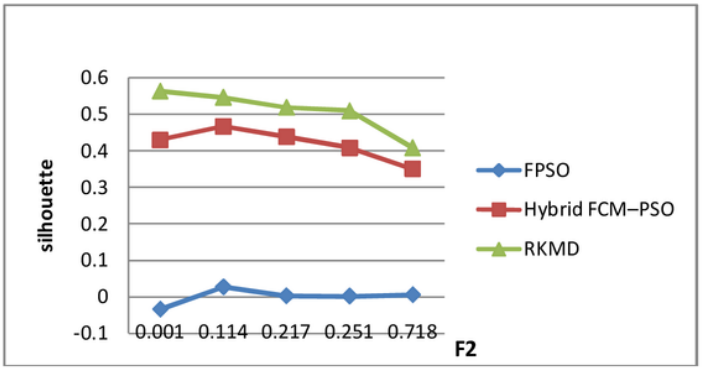



Figure 8. The performance of clustering based on the silhouette index and F2

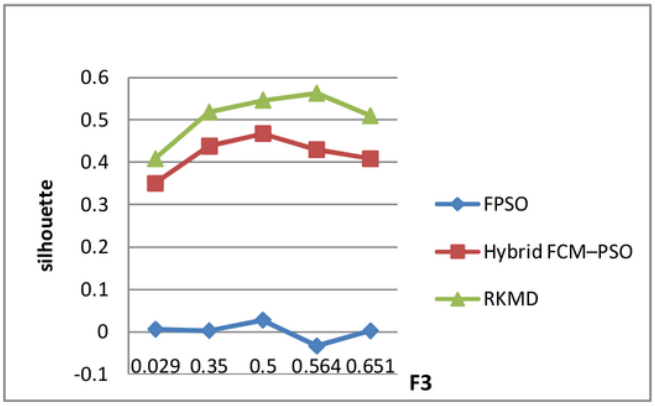



Figure 9. The performance of clustering based on the silhouette index and F3

Figure 8 and Figure 9 indicate that the improvement in the clustering performance is also due to the use of the boundary region. In RKMD, Eq. 12 can be used to define the vague area as well as to reduce the outlier/noise problem. The capability to reduce the outlier problem enables the algorithm to appropriately separate the crisp cluster (lower approximation). Logically, a well-separated lower approximation directly affects the ability to address the uncertainty problem in an overlap area, i.e., the calculation of rough membership. In summary, the use of the hybrid approach in RKMD results better performance than that achieved by the other two algorithms. The clustering performances come from the advantages of RKMD characteristic. We highlight three advantages here:

(1) Initial seed computation relies on the distant concept among the seed. Initial seed is taken from the distant object which is measured by using indiscernibility computation. Since this computation is performed by using binary computation, it requires discretization that is a binary discretization. Moreover it is able to sufficiently select objects as initial seed especially with respect to homogeneity problem. This approach is capable to separate and to restrict area, and to select objects using the discernibility degree. The performance of initial seed computation is indicated by the homogeneity of cluster.

(2) The boundary region defined by RKMD indicates whether an object is vague. This characteristic is useful in some cases as membership does not significantly differentiate objects (for example, indicating which object is closer to the centroid). Therefore, the membership value of the object will be one *(+1)*, or a crisp cluster, whereas in the FPSO and Hybrid FCM–PSO algorithms, all of the objects are viewed as uncertain objects.

This characteristic is also advantageous when the algorithm must avoid coincident clusters. The algorithm's advantage comes from Eq.12. By reducing $\omega_{bnd}$ in our experiment (we use $\omega_{bnd} = 0.15$), the effect of vague objects and outliers on the centroid calculation is decreased.

(3) Based on Eqs. 18 and Eq. 19, the rough membership indicates the belonging of the object. The foundation of the rough membership calculations is granular computation where the lower approximation of RKMD is used as the cluster prototype r. Therefore, the closeness of the object represented by the rough membership is better than by fuzzy clustering since the similarity of the objects is measured based on similarity of the granular knowledge within the set.

## 6. Conclusions

The overlap clustering problem has been studied primarily as it relates to the optimization of membership in fuzzy clustering, which represents the partitioning of overlap clustering. We suggest to process vague objects by using an appropriate method via rough set theory. The main contribution of this work to clustering analysis is its novel approach to overlap clustering based on rough membership computation. This approach is a hybrid of rough k-means and discernibility rough set theory. This hybrid method incorporates the merits of both methods. RKM is performed to find the crisp cluster, which is applied as the foundation to calculate the degree of membership in the overlap cluster. Using our method, clustering performance is improved without increasing the computation time. Additionally, because the method is designed based on discernibility RST, it has advantages in the applications for overlapping classes or naturally occurring partial memberships in the object data.

More efficient use of the information contained in the fuzzy membership function has also been proven by comparing the proposed method using three (3) validation indexes. All of the index values calculated show that the proposed method outperforms with the two other algorithms to which it was compared in the optimization overlap clustering.


### Acknowledgments

I would like to acknowledge Atma Jaya University in Yogyakarta, Indonesia, and UKM Grant No.UKM-DLP-2011-020 for the financial support of this research project.


## References


[1] A. Ali, G. Karmakar, L. Dooley: Review on Fuzzy Clustering Algorithms. J. of Advanced Computations 2(3), 169–181 (2008)

[2] A. Banerjee , C. Krumpelman, J. Ghosh , S. Basu, R.J. Mooney : Model-based overlapping clustering. In: Proceedings of the 11th ACM SIGKDD international conference on knowledge discovery and data mining, Chicago, pp 532–537, 2005

[3] D. Arthur, and S. Vassilvitskii : k-means++: The advantages of careful seeding. In: Proc. ACM-SIAM Symp. Discrete Algorithms (2007)

[4] D.B. Setyohadi, A.A. Bakar, and Z.A. Othman : An Improved Rough Clustering Using Discernibility Based Initial Seed Computation, L. Cao, J. Zhong, and Y. Feng (Eds.): ADMA 2010, Part I, LNCS 6440, pp. 161–168, 2010. © Springer-Verlag Berlin Heidelberg 2010

[5] D.M. Tsai and C. C. Lin : Fuzzy C-means based clustering for linearly and nonlinearly separable data:, Pattern Recognition, 44(2011) 1750-1760.

[6] D. Lee and J. Lee, "Dynamic Dissimilarity Measure for Support Based Clustering," IEEE Trans. Knowledge and Data Eng., vol. 22, no. 6, pp. 900-905, June 2010.

[7] D. Malyszko, J. Stepaniuk, Rough entropy based k-means clustering, in: Rough Sets, Fuzzy Sets, Data Mining and Granular Computing, vol. 5908 of LNCS, 2009, pp. 406–413.

[8] D. Miao, M. Chen, Z. Wei, Q. Duan : A Reasonable Rough Approximation of Clustering Web Users. LNCS Springer, Heidelberg, vol. 4845, pp. 428–442. (2007)

[9] F. Yang, T.Sun, C. Zhang, C. : An efficient hybrid data clustering method based on K-harmonic means, and particle swarm optimization. Expert Systems with Applications(36), 9847–9852.

[10] G. Gan, J. Wu, Z. Yang, A genetic fuzzy k-Modes algorithm for clustering categorical data, Expert Systems with Applications (36), pp. 1615-1620, (2009).

[11] G. Peters : Some Refinement of K-means Clustering. Pattern Recognition, 39, pp: 1481–1491 (2006).

[12] G. Salvatore , B. Matarazzo , R. Slowinski : Parameterized rough set model using rough membership and Bayesian confirmation measures, International Journal of Approximate Reasoning 49 (2008) 285–300

[13] H. Izakian ,Ajith Abraham: Fuzzy C-means and fuzzy swarm for fuzzy clustering problem. Expert Syst. Appl. 38(3): 1835-1838 (2011)
[14] I.T.R. Yanto., T. Herawan, T., &M. Deris,. Data clustering using variable precision rough set (2011). Intelligent Data Analysis, 15(4), 465-482
[15] J.C. Bezdek : Pattern Recognition with Fuzzy Objective Function, Plenum Press, New York, 1981.
[16] J.C. Dunn : A fuzzy relative of the ISODATA process and its use in detecting compact, well-separated clusters, J. Cybernetics 3, 32-57 (1974).
[17] J. Chiang and P. Hao : A new kernel-based fuzzy clustering approach: Support vector clustering with cell growing, IEEE Trans. Fuzzy Syst., vol. 11, no. 4, pp.518 -527 2003
[18] J. Luengo, F. Herrera : Domains of competence of fuzzy rule based classification systems with data complexity measures: a case of study using a fuzzy hybrid genetic based machine learning method Fuzzy Sets and Systems, 161 (2010), pp. 3–19
[19] K.B. Zhang, M. A. Orgun and K. Zhang, "Enhanced Visual Separation of Clusters by M-mapping to Facilitate Cluster Analysis", Proceedings of 9th International Conference series on Visual Information Systems (VISUAL 2007), June 28-29, 2007, Shanghai, China, Lecture Notes in Computer, Volume 4781, Springer Press, pp. 288-300 (2007)
[20] L. Zhu, Fu. Chung, and S. Wang : Generalized Fuzzy C-Means Clustering Algorithm With Improved Fuzzy Partitions IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART B: CYBERNETICS, VOL. 39, NO. 3, JUNE 2009
[21] M. Basu and T. K. Ho : Data Complexity in Pattern Recognition (Advanced Information and Knowledge Processing). Secaucus, NJ,USA: Springer-Verlag New York, Inc., 2006.
[22] M. Chen, D. Miao : Interval set clustering, Expert Systems with Applications: Volume 38 Issue 4, April, 2011
[23] M.H. Wang, Y.F. Tseng, H.C. Chen, K.H. Chao: A novel clustering algorithm based on the extension theory and genetic algorithm. Expert Systems With Applications 36(4), 8269–8276 (2009)
[24] M.K. Ng , M.J. Li, Z.X. Huang, Z.Y. He : On the impact of dissimilarity measure in-Modes clustering algorithm, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (3) (2007) 503–507.
[25] M. Tushir, S. Srivastava : A new Kernelized hybrid c-mean clustering model with optimized parameters. Applied Soft Computing 10 (2010):381-389
[26] N. Taher, A. Babak : An efficient hybrid approach based on PSO, ACO and k-means for cluster analysis, Applied Soft Computing 10 (2010) 183–197
[27] N.R. Pal, K. Pal, J.M. Keller, J.C. Bezdek : A possibilistic fuzzy c-means clustering algorithm, IEEE Transactions on Fuzzy Systems 13 (4) (2005) 517–530.
[28] Pawlak : Rough Sets: Theoretical Aspects of Reasoning about Data, System Theory, Knowledge Engineering and Problem Solving, vol. 9, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1991
[29] P. Lingras, and C. West. Interval set clustering of Web users with rough k-means. Journal of Intelligent Information Systems, 23:5–16, 2004.
[30] P. Maji, S.K. Pal: RFCM: a hybrid clustering algorithm using rough and fuzzy sets FundamentaInformaticae, 79 (2007), pp. 1–22
[31] R. Thangaraj, M. Pant, A. Abraham, P. Bouvry : Particle swarm optimization: hybridization perspectives and experimental illustrations Applied Mathematics and Computation, 217 (12) (2011), pp. 5208–5226
[32] S. Hirano, S. Tsumoto, Indiscernibility-based clustering : Rough clustering, International Fuzzy Systems Association World Congress, LNCS Springer-Verlag, Heidelberg (2003), pp. 378–386Z.
[33] S. Mitra , W. Pedrycz, B. Barman : Shadowed C-Means: Integrating fuzzy and rough clustering Pattern Recognition, 43 (2010), pp. 1282–1291
[34] S. Singh, L. Dey : A new customized document categorization scheme using rough membership, Applied Soft Computing, 5 (4) (2005), pp. 373–390
[35] T. A. Runkler, C. Katz, :Fuzzy clustering by particle swarm optimization. In 2006 IEEE international conference on fuzzy systems (pp. 601–608). Canada.
[36] T.K. Ho, M. Basu, Complexity Measures of Supervised Classification Problems, IEEE Transactions on Pattern Analysis and Machine Intelligence, 24 , 3, March 2002, 289-300.

# ccp1

GRADEMARK REPORT

FINAL GRADE

/0

GENERAL COMMENTS

**Instructor**

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10

PAGE 11

PAGE 12

PAGE 13

PAGE 14

PAGE 15

PAGE 16

PAGE 17

PAGE 18

PAGE 19

PAGE 20

PAGE 21

PAGE 22

PAGE 23

PAGE 24

PAGE 25

PAGE 26

PAGE 27

# ccp1

ORIGINALITY REPORT

| 12% | 6% | 5% | 8% |
|---|---|---|---|
| SIMILARITY INDEX | INTERNET SOURCES | PUBLICATIONS | STUDENT PAPERS |

PRIMARY SOURCES

| | | |
|---|---|---|
| 1 | Submitted to Universiti Putra Malaysia<br>Student Paper | 6% |
| 2 | content.iospress.com<br>Internet Source | 3% |
| 3 | Djoko Budiyanto Setyohadi. "An Improved Rough Clustering Using Discernibility Based Initial Seed Computation", Lecture Notes in Computer Science, 2010<br>Publication | 1% |
| 4 | www.softcomputing.net<br>Internet Source | 1% |
| 5 | Peters, Georg, Fernando Crespo, Pawan Lingras, and Richard Weber. "Soft clustering – Fuzzy and rough approaches and their extensions and derivatives", International Journal of Approximate Reasoning, 2012.<br>Publication | 1% |
| 6 | comp.mq.edu.au<br>Internet Source | 1% |

| | | | |
|---|---|---|---|
| Exclude quotes | Off | Exclude matches | < 1% |
| Exclude bibliography | Off | | |