Variance based Three-way Clustering Approaches for Handling Overlapping Clustering

Mohammad Khan Afridi[a], Nouman Azam[a,*], JingTao Yao[b]

[a] *National University of Computer and Emerging Sciences, Pakistan*
[b] *Department of Computer Science, University of Regina, Regina, SK, S4S 0A2, Canada*

Abstract

The conventional clustering approaches are not very effective in dealing with clusters having overlapping regions. The three-way clustering (3WC) is an effective and promising approach in this regards. A key issue in 3WC is the determination of thresholds which plays a crucial and important role in accurate estimation of the overlapping region. In this article, we propose different variance based criteria for determining the thresholds. In particular, we examine the variance or spread in evaluation function values of objects contained in the three regions obtained with 3WC of objects. An algorithm called 3WC-OR is introduced that considers the optimization of the proposed criteria for determining effective thresholds by incorporating approaches such as genetic algorithms and game-theoretic rough sets. Experimental results on five UCI datasets indicate that the proposed algorithm significantly improves results on datasets with overlapping clusters and provide comparable results on datasets with non-overlapping clusters.

1. Introduction

In many applications, clusters may have unclear and vague boundaries and have overlapping region [34]. The conventional hard clustering approaches are not suitable in such applications. In particular, they assign each object to ex-

*Corresponding author
Email address: nouman.azam@nu.edu.pk (Nouman Azam)

actly one cluster and results in non-overlapping and disjoint clusters. Dealing with overlapping clusters is an important data analysis task with many application areas including categorization of products into multiple categories, predicting diseases with common symptoms, social network users belong to multiple groups and multi-label assignment to textual documents [2, 21, 35, 47].

Many approaches and methods have been introduced for handling of overlapping clustering. One class of these approaches is based on the extensions of well known hard clustering methods such as k-means, k-centroids and k-medoids [5, 8, 17, 20, 25]. The general idea in these approaches is to use some threshold values on the results obtained with hard clustering approaches to produce overlapping clusters [9]. An object can therefore belong to all those clusters for which its association satisfies the threshold values. Another class of approaches is based on soft clustering [11, 30, 32, 48]. These approaches make multiple cluster assignment by considering the use of suitable thresholds on objects association with different clusters. Despite the differences, the choice of a suitable threshold remains an unsolved issue in the two classes or types of approaches [9, 33]. There are other approaches including star, estar, suffix tree based and connected graph based iterative scan [3, 13, 14, 46]. These approaches generally involve very high computation and are therefore not very feasible [28].

A more recent proposal for dealing with overlapping clustering is based on three-way clustering (3WC) framework which was proposed by Yu et al. [41, 42, 43, 44]. In contrast to hard clustering which uses a single set to represent each cluster, the 3WC uses two sets to represent each cluster. The two sets are further processed to define three types of relationships between an object and a cluster, namely, belongs to, not belongs to and partially belongs to a cluster. Objects with relationship belongs to are definitely part of the cluster, and only belong to that cluster while objects with relationship not belongs to are definitely not part of the cluster. The objects with relationship partially belongs to may possibly be part of the cluster and may potentially belong to some other clusters. Such objects may therefore be considered as part of the overlapping region. The essential ideas of 3WC were also considered by other studies including

Lingras and Peters [19], Pedrycz [27], Lingras and West [20], Campagner and Ciucci [7], Yao et al. [38] and Zhou et al. [49] under the names of rough clustering, interval set clustering and shadowed sets clustering. Like 3WC, the interval and rough set clustering also make use of two sets for approximating the clusters, however, they impose different conditions on the two sets corresponding to a cluster [7]. The shadowed sets based clustering makes use of thresholds on fuzzy membership values to achieve three-way clustering [27].

The three types of relationships in 3WC and the resulting overlapping regions are critically defined based on a pair of (α, β) thresholds. In particular, the objects whose association with the cluster is above the threshold α are considered to have belongs to relationship while objects whose association with the cluster is below the threshold β are considered to have not belongs to relationship. The objects whose association with the cluster falls between the two thresholds are considered to have partially belongs to relationship. Automated determination of optimal values of these thresholds is a crucial research challenge in this context.

In this paper, we extend the existing studies of 3WC by considering this issue while estimating effective overlapping regions of clusters. Variance is a typical measure that is frequently used in evaluation of clusters and cluster formation. In particular, we propose variance based 3WC approaches that make use of optimization criteria for automatically determining the thresholds. Two types of variances in evaluation function values of objects are considered for this purpose, namely, within region variance and between region variance. The within region variance measures the spread of evaluation function values of objects within a certain region. Generally speaking, one would like to reduce the within region variance which results in more similar evaluation values for objects within a certain region, resulting in better grouping of objects in that region. On the other hand, the between region variance measures the spread in region based mean of evaluation function values compared to global mean of evaluation function values. One would also like to maximize this which results in well separated and distinguishable regions in evaluation function values. We approach

the determination of thresholds based on criteria that consider the optimization of the two types of variances. An algorithm called 3WC-OR is introduced that considers the optimization of the proposed criteria by consider iterative learning approaches such as genetic algorithms and game-theoretic rough sets. Experimental results on five UCI datasets suggest that the 3WC-OR algorithm lead to better estimation of the overlapping region compared to other three-way and similar approaches. In particular, it improves accuracy by up to 4% for datasets with overlapping clusters while having similar accuracy on datasets with non-overlapping clusters.

2. Three-way Clustering for Estimating Overlapping Region of Clusters

In this section, we first provide the relevant background by discussing the key notions of 3WC. Later, in the same section, we highlight the importance of thresholds determination for accurate estimation of the overlapping region of clusters which sets the motivation for the present study.

2.1. Basics of Three-way Clustering

In this section, we provide the basic notions of three-way clustering. Consider a set of objects $U = \{o_1, o_2, o_3, ...\}$ and a family of clusters $C = \{c_1, c_2, c_3, ...\}$. We use two sets for representing each cluster c_k, i.e.,$c_k = \{In(c_k), Pt(c_k)\}$, where, $In(c_k)$ and $Pt(c_k) \subset U$. The three regions of a cluster are defined using the two sets as follows [45, 41].

$$Inside(c_k) = In(c_k), \tag{1}$$

$$Partial(c_k) = Pt(c_k), \tag{2}$$

$$Outside(c_k) = U - In(c_k) - Pt(c_k). \tag{3}$$

The $Inside(c_k)$ contains objects that definitely belongs to the cluster c_k and only belong to c_k, the $Outside(c_k)$ contains objects that does not belong to the cluster c_k and the $Partial(c_k)$ contains objects which may belong to the cluster c_k and possibly to some other clusters. The objects in the partial region are

therefore potentially overlapping objects among the clusters. The three regions are reflective of three types of relationships between an object and a cluster. In particular, the objects in the $Inside(c_k)$ and $Outside(c_k)$ are reflective of belongs to and not belongs to relationships while the objects in the $Partial(c_k)$ are reflective of partially belongs to relationship.

To obtain the three regions, the 3WC adopts the three-way decisions formulation based on an evaluation function and a pair of thresholds [39]. Considering $e(c_k, o_i)$ to be an evaluation function that quantifies the relationship between an object o_i and cluster c_k. By considering the thresholds (α, β), the three regions are defined as,

$$Inside_{(\alpha,\ \beta)}(c_k) = \{o_i \in U \mid e(c_k, o_i) \geq \alpha\}, \tag{4}$$

$$Partial_{(\alpha,\ \beta)}(c_k) = \{o_i \in U \mid \beta < e(c_k, o_i) < \alpha\}, \tag{5}$$

$$Outside_{(\alpha,\ \beta)}(c_k) = \{o_i \in U \mid e(c_k, o_i) \leq \beta\}. \tag{6}$$

An object o_i belongs to the $Inside_{(\alpha,\beta)}(c_k)$ region when its evaluation function value is at or above threshold α. Similarly, an object o_i belongs to the $Outside_{(\alpha,\ \beta)}(c_k)$ region when its evaluation function value is at or below threshold β. Finally, an object belongs to the $Partial(c_k)$ region if its evaluation function value is lesser than α but greater than β.

The configuration of thresholds (α, β) controls the inclusions of objects into different regions and therefore plays a key role in the estimation of overlapping region. We further elaborate on this in the next section.

2.2. Importance of Thresholds Determination for Overlapping Region

In this section, we highlight the importance of determining suitable thresholds of 3WC for accurate estimation of clusters overlapping region. Consider Table 1 that contains information about 42 objects in its rows, represented as o_1, o_2, ... , o_{42}. The objects are described based on 4 attributes, represented as A_1, A_2, A_3 and A_4. We assume that there are two clusters defined by the sets $c_1 = \{o_1, ..., o_{26}\}$ and $c_2 = \{o_{18}, ..., o_{42}\}$. The objects that are present in

Table 1: Example Dataset

	A_1	A_2	A_3	A_4		A_1	A_2	A_3	A_4		A_1	A_2	A_3	A_4
o_1	7	3.2	4.7	1.4	o_{15}	5.8	2.6	4.1	1	o_{29}	7.3	2.9	6.3	1.8
o_2	6.4	3.2	4.5	1.5	o_{16}	6.4	3.1	4.5	1.5	o_{30}	5.8	2.8	5.1	2.4
o_3	5.2	2.6	4.4	1.6	o_{17}	5.6	2.5	3.9	1.1	o_{31}	7.1	3.5	6.1	2.5
o_4	5.6	2.3	4.1	1.3	o_{18}	6.3	3	4.9	2	o_{32}	6.7	3	5.5	1.9
o_5	6.5	2.8	4.6	1.5	o_{19}	4.9	2.4	3.3	1	o_{33}	7.7	3.8	6.7	2.2
o_6	5.9	2.8	4.5	1.3	o_{20}	6.3	3	4.9	1.8	o_{34}	7.7	2.6	6.9	2.3
o_7	6.1	3	4.6	1.4	o_{21}	5.6	3	4.5	1.5	o_{35}	6.5	3	5.5	1.8
o_8	5	2.4	3.3	1.1	o_{22}	6.4	2.7	5.3	1.9	o_{36}	6.9	3.2	5.7	2.3
o_9	5.9	3	4.2	1.5	o_{23}	5.7	2.5	5	2	o_{37}	7.2	3.6	6.1	2.5
o_{10}	6	2.2	4	1	o_{24}	6.7	2.5	5.8	1.8	o_{38}	7.7	2.8	6.7	2
o_{11}	6.1	2.9	4.7	1.4	o_{25}	6.1	2.3	5	1.5	o_{39}	6.9	3	5.1	1.7
o_{12}	5.6	2.9	3.6	1.3	o_{26}	6.5	3.2	5.1	2	o_{40}	6.7	3.3	5.7	2.1
o_{13}	6.7	3.1	4.4	1.4	o_{27}	6.3	2.9	4.5	1.3	o_{41}	7.2	3.2	6	1.8
o_{14}	5.2	2.6	4.4	1.5	o_{28}	6.8	3	5.5	2.1	o_{42}	7.1	3	6.1	1.9

the both the clusters forms the overlapping region which is given by the intersection $c_1 \cap c_2 = \{o_{18}, ..., o_{26}\}$. Moreover, we consider an object to be in the overlapping region if it does not have a belongs to relationship with any cluster or equivalently it is not in the inside of any cluster.

To compute the 3WC of objects, we need to determine evaluation function values for all the objects. We use the evaluation function which considers the number of neighbors of o_i belonging to c_k, i.e.,

$$e(c_k, o_i) = \frac{\text{Number of neighbors of } o_i \text{ belonging to } c_k}{\text{Total neighbors of } o_i}. \tag{7}$$

Equation (7) highlights the relationship between an object o_i and cluster c_k based on the relative number of neighbors of each object o_i in cluster c_k. The more the neighbors, the stronger the relationship. A slightly different evaluation functions are used in the previous literature [33, 41]. The neighbors of an object o_i are computed by measuring the distances of all the objects from o_i based on a certain distance metric, such as Euclidean distance, and then sort these distances

in an ascending order. For instance, the Euclidean distances of all objects in U from o_1 are $d(o_1, o_2) = 0.64$, $d(o_1, o_3) = 1.93$, ..., $d(o_1, o_{42}) = 1.5$. Sorting these distances, results in the following seven nearest neighbors of o_1 i.e., o_{13}, o_{39}, o_2, o_{16}, o_5, o_{27} and o_{20}. The information of neighbors is next used to determine evaluation function values. For instance, evaluation function value for object o_1 and cluster c_1 based on its seven nearest neighbors is,

$$e(c_1, o_1) = \frac{\text{Number of } o_1 \text{ neighbors belong to } c_1}{\text{Total neighbors of } o_1} = 5/7 = 0.71, \qquad (8)$$

which means that 71% neighbors of object o_1 belong to cluster c_1.

Once the evaluation function values are computed for all objects, we can compute 3WC of objects based on some thresholds using Equations (4) - (6). For instance, for thresholds (α, β) = (1, 0), the inside regions for two clusters c_1 are given by,

$$Inside_{(1,\ 0)}(c_1) = \{o_1, o_2, o_5, o_6, o_7, o_9, o_{10}, o_{11}, o_{12}, o_{13}, o_{15}, o_{16}, o_{27}\}, \quad (9)$$

$$Inside_{(1,\ 0)}(c_2) = \{o_{29}, o_{30}, o_{31}, o_{33}, o_{34}, o_{36}, o_{37}, o_{38}, o_{40}, o_{41}, o_{42}\}, \quad (10)$$

The consideration of different threshold settings produce different regions and therefore leads to different overlapping regions. For example, consider the two extreme cases that are associated with thresholds setting of $(\alpha, \beta) = (1, 0)$ and $(\alpha, \beta) = (0.5, 0.5)$. In case of $(\alpha, \beta) = (1, 0)$, there are strict conditions for inclusion in the inside and outside regions which are only satisfied by a few objects and therefore many objects ends up in the partial region. For the $(\alpha, \beta) = (0.5, 0.5)$, there are relax and lose conditions for inclusion in the inside and outside region which are easily satisfied by majority of the objects and therefore has least objects in the partial region. In case of (α, β) = (1, 0), we have 24 objects in the inside of the two clusters based on Equations (9) - (10), i.e., $|Inside_{(1,\ 0)}(c_1)| + |Inside_{(1,\ 0)}(c_2)| = 13 + 11 = 24$. This means that the remaining 42-24 = 18 objects forms the overlapping region. The overlapping region in this case is almost twice in size compared to the original clustering assignment in Table 1, i.e., 18 versus 9. On the other hand, in case of (α, β) = (0.5, 0.5). The inside regions for two clusters in this case contains all 42

objects, i.e., $|Inside_{(0.5,\ 0.5)}(c_1)| + |Inside_{(0.5,\ 0.5)}(c_2)| = 23 + 19 = 42$ which means that none of the objects belong to the overlapping region. From above discussion, we may note that in one extreme setting of thresholds, we have an over estimation of the overlapping region and in another extreme case we have an underestimation of the overlapping region. This intuitively suggests that a more reasonable estimate of the overlapping region is be between these two extreme settings of thresholds.

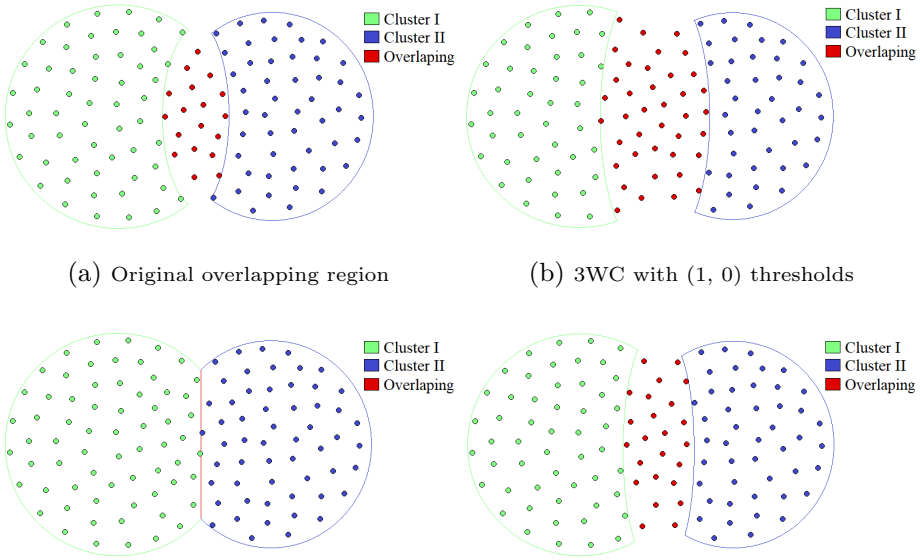
Now let us look at moderate setting of thresholds between these two extreme cases, such as, (α, β)= (0.75, 0.25). There are 34 objects in the inside of the two clusters, i.e., $|Inside_{(0.75,\ 0.25)}(c_1)| + |Inside_{(0.75,\ 0.25)}(c_2)| = 19 + 15 = 34$ which suggest that the remaining 8 objects form the overlapping region. This is comparatively more reasonable and more accurate estimate of the overlapping region which is neither overestimated nor underestimated.

There are two main observations from the above example. First, in the 3WC, the threshold values are critically important for accurate and effective estimation of the overlapping clusters region. Second, effective values for the thresholds lie between the two extreme configuration of the thresholds settings, i.e., $(\alpha, \beta) = (1, 0)$ and $(0.5, 0.5)$. In the light of these two observations, the determination of effective thresholds becomes a key research issue and challenge. We use visual representation to further elaborate on this in the next section.

2.3. A Visual Representation of Thresholds and Overlapping Region Relationship

This section provides a visual representation for highlighting the relationship between thresholds and its impact on the estimation of the clusters overlapping region. Figure 1 (a) - (d) is be used for this purpose. There are two clusters in each of the sub figures. The objects in the two clusters are represented with small green and brown circles while the objects in the overlapping region are represented with red circles.

Figure 1 (a) depicts the actual overlapping of clusters. Figure 1 (b) represents the overlapping region for the thresholds of $(\alpha, \beta) = (1, 0)$. An object is



(a) Original overlapping region (b) 3WC with (1, 0) thresholds

(c) 3WC with (0.5, 0.5) thresholds (d) 3WC with moderate thresholds

Figure 1: Comparison between clusters with their evaluation function values

included in the non-overlapping region only when its evaluation function value has a value of 1 for any cluster. A significant number of objects does not meet this strict requirement and has evaluation value of lesser than 1. This results in large overlapping region. Figure 1 (c) represents clustering with thresholds of $(\alpha, \beta) = (0.5, 0.5)$. An object is included in the non-overlapping region when its evaluation function value for a cluster is greater than or equal to 0.5. Since this condition is easily met for all the objects, therefore, it results in empty overlapping region. Figure 1 (d) represents clustering with thresholds between the two extreme thresholds. We may notice that the overlapping region in this case is more appropriately representing the actual overlapping region. In contrast to $(\alpha, \beta) = (1, 0)$, it has lesser objects in the overlapping region and in contrast to $(\alpha, \beta) = (0.5, 0.5)$ it has more objects in the overlapping region.

A key issue is that the thresholds control the overlapping region and one simply cannot fix the thresholds to some arbitrary value and expect an optimum estimation of the overlapping region. We need to formulate a method in order

to obtain thresholds between the two extreme cases namely, $(\alpha, \beta) = (1, 0)$ and $(\alpha, \beta) = (0.5, 0.5)$. We propose variances based criteria for this purpose in this article.

3. Variance based Three-way Clustering

We introduce variance based criteria for determining 3WC thresholds in this section. In particular, we investigate how the changes of thresholds affect the variance in evaluation function values for the objects in the three regions. The statistical measures has recently gained some attention for determining thresholds of three-way decisions [12, 40]. However, unlike the existing studies, we look at the statistical measures from the viewpoint of optimization.

3.1. Basic Formulation

Considering a threshold pair (α, β), which are used to create three regions of $Inside(c)$, $Partial(c)$ and $Outside(c)$ regions, corresponding to a cluster. The mean of evaluation function values for the objects included in the three regions are computed as,

$$\mu_{Inside_{(\alpha,\,\beta)}(c)} = \frac{\sum_{\forall o_i \in Inside_{(\alpha,\,\beta)}(c)} e\big(c, o_i\big)}{|Inside_{(\alpha,\,\beta)}(c)|}, \tag{11}$$

$$\mu_{Partial_{(\alpha,\,\beta)}(c)} = \frac{\sum_{\forall o_i \in Partial_{(\alpha,\,\beta)}(c)} e\big(c, o_i\big)}{|Partial_{(\alpha,\,\beta)}(c)|}, \tag{12}$$

$$\mu_{Outside_{(\alpha,\,\beta)}(c)} = \frac{\sum_{\forall o_i \in Outside_{(\alpha,\,\beta)}(c)} e\big(c, o_i\big)}{|Outside_{(\alpha,\,\beta)}(c)|}. \tag{13}$$

The mean of each region is the sum of evaluation function values of all the objects in that regions divided by the total objects in that region. For instance, $\mu_{Inside_{(\alpha,\,\beta)}(c)} = 0.91$ will mean that for all the objects, in the $Inside(c)$ region of cluster c, the mean evaluation function value is 0.91. The overall mean of the evaluation function values in the entire dataset is computed as,

$$\mu = \frac{\sum_{\forall o_i \in C} e(c, o_i)}{|U|}. \tag{14}$$

The variance in evaluation function values for a region represents the spread of evaluation function values within that region. The variance of the three regions are computed as,

$$\sigma^2_{Inside_{(\alpha,\,\beta)}(c)} = \sum_{\forall o_i \in Inside_{(\alpha,\,\beta)}(c)} \Big(e(c,o_i) - \mu_{Inside_{(\alpha,\,\beta)}(c)}\Big)^2, \quad (15)$$

$$\sigma^2_{Partial_{(\alpha,\,\beta)}(c)} = \sum_{\forall o_i \in Partial_{(\alpha,\,\beta)}(c)} \Big(e(c,o_i) - \mu_{Partial_{(\alpha,\,\beta)}(c)}\Big)^2, \quad (16)$$

$$\sigma^2_{Outside_{(\alpha,\,\beta)}(c)} = \sum_{\forall o_i \in Outside_{(\alpha,\,\beta)}(c)} \Big(e(c,o_i) - \mu_{Outside_{(\alpha,\,\beta)}(c)}\Big)^2. \quad (17)$$

Ideally, we would like to have minimum possible variance of evaluation function values in each region. By reducing the within region variance, the evaluation values of the objects within a region will be more similar to each other, resulting in better grouping of objects. In the next section, we explain how the notion of regional means and variances of evaluation function values can be used to define criteria for optimizing thresholds.

3.2. Variance based Optimization Criteria for Determining 3WC Thresholds

The optimization based determination of thresholds for three-way decisions has been adopted by many researchers [10]. In particular, the determination of thresholds is formulated as optimization of distinct regional properties such as accuracy, generality, risk, variance and others [4]. Let $Q_{(\alpha,\,\beta)}(c)$ be a measure representing a certain quality related aspect of the three regions. In most instances, the quality measure $Q_{(\alpha,\,\beta)}(c)$ is a combination of the $Q_{\mathrm{POS}_{(\alpha,\,\beta)}}(c)$, $Q_{\mathrm{BND}_{(\alpha,\,\beta)}}(c)$ and $Q_{\mathrm{NEG}_{(\alpha,\,\beta)}}(c)$, representing some qualitative aspect of the POS, BND and NEG regions, respectively. The overall quality of the three regions is expressed as,

$$Q_{(\alpha,\,\beta)}(c) = w_1\, Q_{\mathrm{POS}_{(\alpha,\,\beta)}}(c) + w_3\, Q_{\mathrm{NEG}_{(\alpha,\,\beta)}}(c) + w_2\, Q_{\mathrm{BND}_{(\alpha,\,\beta)}}(c), \quad (18)$$

where w_1, w_2 and w_3 are weights associated with the three regions. More formally, the determination of thresholds is approached as the following opti-

mization problem [10],

$$\arg\min_{(\alpha,\,\beta)} Q_{(\alpha,\,\beta)}(c) \;\text{ or }\; \arg\max_{(\alpha,\,\beta)} Q_{(\alpha,\,\beta)}(c). \tag{19}$$

In the context of 3WC, Equation (18) is represented as,

$$Q_{(\alpha,\,\beta)}(c) = w_1\, Q_{Inside_{(\alpha,\,\beta)}(c)} + w_2\, Q_{Partial_{(\alpha,\,\beta)}(c)} + w_3\, Q_{Outside_{(\alpha,\,\beta)}(c)}. \tag{20}$$

We examine variance based criteria for determining the thresholds. The first criterion we examine is the within region variance. It is computed as the weighted sum of the individual region variances (Equations (15) - (17)). The regional variances of the three regions are,

$$\sigma^2_{W_{Inside_{(\alpha,\,\beta)}(c)}} = \sigma^2_{Inside_{(\alpha,\,\beta)}(c)}, \tag{21}$$

$$\sigma^2_{W_{Partial_{(\alpha,\,\beta)}(c)}} = \sigma^2_{Partial_{(\alpha,\,\beta)}(c)}, \tag{22}$$

$$\sigma^2_{W_{Outside_{(\alpha,\,\beta)}(c)}} = \sigma^2_{Outside_{(\alpha,\,\beta)}(c)}, \tag{23}$$

where the additional notations are being added for the sake of being consistent with the notations introduced in Equation (20). Next, the overall region variance is defined as the weighted sum of the three within region variances as follows,

$$\sigma^2_{W_{(\alpha,\,\beta)}(c)} = w_1 \times \sigma^2_{W_{Inside_{(\alpha,\,\beta)}(c)}} + w_2 \times \sigma^2_{W_{Partial_{(\alpha,\,\beta)}(c)}} + w_3 \times \sigma^2_{W_{Outside_{(\alpha,\,\beta)}(c)}}. \tag{24}$$

Equation (24) measures the overall spread of the evaluation function values within each region. Minimizing this will lead to regions containing objects having lesser variation in their evaluation function values. Since the evaluation function values represent the relationship between an object and a cluster. Therefore, this will intuitively mean that the objects in a certain region exhibits very similar relationship with the cluster. This leads to the following optimization criterion,

$$arg\min_{(\alpha,\,\beta)} \sigma^2_{W_{(\alpha,\,\beta)}(c)}. \tag{25}$$

where $u_{Inside_{(\alpha,\beta)}}(c)$, $u_{Partial_{(\alpha,\beta)}}(c)$, and $u_{Outside_{(\alpha,\beta)}}(c)$, are defined in Equations (11) - (13). Next, we define another criterion called the between region variance. The between region variance is defined as the difference of regional

mean of evaluation function values and global mean of evaluation function values. It is computed as follow,

$$\sigma^2_{B_{Inside_{(\alpha,\,\beta)}(c)}} = \big(\mu_{Inside_{(\alpha,\,\beta)}(c)} - \mu\big)^2, \tag{26}$$

$$\sigma^2_{B_{Partial_{(\alpha,\,\beta)}(c)}} = \big(\mu_{Partial_{(\alpha,\,\beta)}(c)} - \mu\big)^2, \tag{27}$$

$$\sigma^2_{B_{Outside_{(\alpha,\,\beta)}(c)}} = \big(\mu_{Outside_{(\alpha,\,\beta)}(c)} - \mu\big)^2. \tag{28}$$

Next, the overall between region variance is computed as a the average sum of the between region variances as follows,

$$\sigma^2_{B_{(\alpha,\,\beta)}(c)} = w_1 \times \sigma^2_{B_{Inside_{(\alpha,\,\beta)}(c)}} + w_2 \times \sigma^2_{B_{Partial_{(\alpha,\,\beta)}(c)}} + w_3 \times \sigma^2_{B_{Outside_{(\alpha,\,\beta)}(c)}}. \tag{29}$$

Equation (29) measures the spread or variance of regional means compared to global mean or mean of the entire population. Maximizing this will result in well separated and distinguishable regions in evaluation function values. This leads to the following optimization criterion,

$$arg \max_{(\alpha,\,\beta)} \sigma^2_{B_{(\alpha,\,\beta)}(c)}. \tag{30}$$

The within region variance and between region variance can also be used together for determining thresholds which leads to the following criterion given below,

$$arg \max_{(\alpha,\,\beta)} \frac{\sigma^2_{B_{(\alpha,\,\beta)}(c)}}{\sigma^2_{W_{(\alpha,\,\beta)}(c)}}. \tag{31}$$

Equation (31) is a frequently used discriminant criterion used in discriminant analysis [4]. The underlying rationale in the proposed variance based criteria is that well thresholded regions should be well separated in evaluation function values. In other words thresholds resulting in the best separation of regions in evaluation function values are the optimal threshold values. A learning algorithm may be employed that could search the space of possible thresholds for determining thresholds that satisfy the criteria specified in Equations (25), (30) and (31). We further discuss this in Section 4.

We consider the weights in Equations (24) and (29) as the probabilities of the three regions. The probability of a certain region, for instance, inside region,

is computed as,

$$P(Inside_{(\alpha,\beta)}(C) = |Inside_{(\alpha,\beta)}(C)|/|U|). \quad (32)$$

This means now that the region with high probability or size will contribute more to the overall variance compared to a region with lesser size.

3.3. Estimating Overlapping Regions with Variance Based Criteria

In this section, we demonstrate how variance based criteria introduced in Section 3.2 can be used for effective estimation of the overlapping region by determining suitable thresholds for 3WC 2.2. In Section 2.3, we highlighted that the use of different threshold settings lead to different overlapping regions and we therefore need to have a mechanism which will enable us to choose suitable threshold values. From Section 3.2, we may note the different variance based criteria depends on the (α, β) thresholds. This means that the by changing the thresholds we can expect different values of these criteria and therefore different overlapping regions. We now show how the choice of different threshold values will affect the variance based criteria and the resulting overlapping regions. For this purpose, we compute the within region variance and between region variance for different thresholds settings. The dataset of Table 1 introduced in Section 2.2 will be used for this purpose.

We first compute the two types of variances for some random threshold values, say (α, β) = (0.8, 0.3) for cluster c_1. The three regions corresponding to cluster c_1 based on considered thresholds are,

$$Inside_{(0.8,\ 0.3)}(c_1) = \{o_2, ..., o_{17}, o_{19}, o_{21}, o_{27}\}, \quad (33)$$

$$Partial_{(0.8,\ 0.3)}(c_1) = \{o_1, o_{18}, o_{20}, o_{23}, o_{25}\}, \quad (34)$$

$$Outside_{(0.8,\ 0.3)}(c_1) = \{o_{22}, o_{24}, o_{26}, o_{28}, ..., o_{42}\}. \quad (35)$$

The three regions can be used to compute the mean evaluation function values for all the objects in the three regions of cluster c_1. The mean evaluation function value for a certain region, say $Inside(c_1)$ is computed by taking the

mean of evaluation function values for all objects belonging to that region. This is given by Equation (11), which can be computed as,

$$\begin{aligned}\mu_{Inside_{(0.8,\,0.3)}(c_1)} &= \frac{\sum\limits_{\forall o_i \in Inside_{(0.8,\,0.3)}(c_1)} e\big(c_1, o_i\big)}{|Inside_{(0.8,\,0.3)}(c_1)|}, \\ &= \frac{e(c_1, o_2) + ... + e(c_1, o_{27})}{|o_2, ..., o_{27}|} = 0.91. \end{aligned} \tag{36}$$

where the evaluation function values for a certain object is computed based on Equation (7). In the same way, we can compute the mean for the outside and partial regions of c_1 which for the considered thresholds are $\mu_{Outside_{(0.8,\,0.3)}(c_1)} = 0.06$ $\mu_{Partial_{(0.8,\,0.3)}(c_1)} = 0.66$, respectively.

Next, we compute the variance of the three regions based on the mean of the three regions. To measure the within region variance of all the regions, we use Equations (15) - (17). For instance, the within region variance of the inside region is computed as,

$$\begin{aligned}\sigma^2_{Inside_{(0.8,\,0.3)}(c_1)} &= \sum_{\forall o_i \in Inside_{(0.8,\,0.3)}(c_1)} \Big(e(c_1, o_i) - \mu_{Inside_{(0.8,\,0.3)}(c_1)}\Big)^2, \\ &= (e(c_1, o_2) - \mu_{Inside_{(0.8,\,0.3)}(c_1)})^2 + ... + \\ &\quad (e(c_1, o_{27}) - \mu_{Inside_{(0.8,\,0.3)}(c_1)})^2 = 0.09. \end{aligned} \tag{37}$$

Similarly, we can compute the within region variance for the outside and partial regions which are $\sigma^2_{Outside_{(0.8,\,0.3)}(c_1)} = 0.26$ and $\sigma^2_{Partial_{(0.8,\,0.3)}(c_1)} = 0.025$, respectively. Next, we use Equation (24) to determine the within region variance for the thresholds $(\alpha, \beta) = (0.8, 0.3)$.

$$\begin{aligned}\sigma^2_{W_{(0.8,\,0.3)}}(c_1) &= \sigma^2_{W_{Inside_{(0.8,\,0.3)}(c_1)}} \times w_1 + \sigma^2_{W_{Outside_{(0.8,\,0.3)}(c_1)}} \times w_2 \\ &\quad + \sigma^2_{W_{Partial_{(0.8,\,0.3)}(c_1)}} \times w_3, \\ &= 0.09 \times \frac{19}{42} + 0.26 \times \frac{18}{42} + 0.025 \times \frac{5}{42} = 0.155, \end{aligned} \tag{38}$$

where w_1, w_2 and w_3 are considered as the probabilities of the inside, outside and partial regions, respectively.

Let us now compute the second criterion of between region variance. To compute that we need to determine mean of the evaluation function values

of objects in the whole clustering assignments. Equation (14) can be used to determine the evaluation mean of the 3WC assignments for (α, β) threshold equal to (0.8, 0.3),

$$\begin{aligned} \mu_{(0.8,\,0.3)} &= \frac{\sum\limits_{\forall o_i \in C} e(c, o_i)}{|U|}, \\ &= \frac{e(c_1, o_1) + ... + e(c_1, o_{42})}{42} = 0.52. \end{aligned} \tag{39}$$

Using Equation (29) we determine the between region variance of the clustering assignments for (α, β) = (0.8, 0.3).

$$\begin{aligned} \sigma^2_{B_{(0.8,0.3)}} &= \sigma^2_{B_{Inside_{(0.8,0.3)}(c_1)}} \times w_1 + \sigma^2_{B_{Outside_{(0.8,0.3)}(c_1)}} \times w_2 \\ &\quad + \sigma^2_{B_{Partial_{(0.8,0.3)}(c_1)}} \times w_3, \end{aligned} \tag{40}$$

$$\begin{aligned} &= (0.91 - 0.52)^2 \times \frac{19}{42} + (0.06 - 0.52)^2 \times \frac{18}{42} + (0.66 - 0.52)^2 \times \frac{5}{42}, \\ &= 0.069 + 0.091 + 0.002 = 0.162. \end{aligned} \tag{41}$$

The above demonstration explained how the within and between region variances are determined for a certain threshold values. By considering the domains of possible thresholds that may happen within the data, we can compute the corresponding within and between region variances and then selecting the thresholds with suitable values for these criteria.

Table 2: Effect of different threshold values on between and within region variances

		α			
		1	0.8	0.6	0.5
β	0	(0.145, 0.508)	(0.159, 0.112)	(0.162, 0.115)	(0.16, 0.21)
	0.3	(0.159, 0.176)	(0.16, 0.153)	(0.158, 0.208)	(0.154, 0.319)
	0.5	(0.16, 0.177)	(0.16, 0.152)	(0.16, 0.21)	(0.15, 0.32)

Table 2 shows the values of the two types of variances corresponding to different thresholds values that may happen with in the data. We may note from Table 2 that the two extreme settings of the thresholds, i.e., (1, 0) and

(0.5, 0.5), which do not provide effective estimate of the overlapping regions, also does not provide optimal results for the two criteria. The effective thresholds are between these two cases and may be found by searching the entire table. For instance, the thresholds (1, 0.5) and (0, 0.5) provides the best results for the between region variance and the thresholds (0.8, 0.3) provides effective results for the within region variance. One may also compute the ratio of the two variances to find the optimal thresholds based on the two criteria. Generally, it is not always feasible to list all possible combinations of the thresholds and do an exhaustive search for finding suitable thresholds. A learning algorithm may be introduced for searching of effective thresholds. We discuss this in detail in the next section.

4. Variance based 3WC Algorithm for Estimating Overlapping Region

This section introduces 3WC-OR or 3WC algorithm for estimating overlapping region. The algorithm determines thresholds of 3WC by considering the variance based criteria introduced in Section 3 and its optimization using learning algorithms. The determined thresholds are then used to estimate overlapping region. The algorithm is presented as Algorithm 1.

Algorithm 1. A 3WC Algorithm for Overlapping Region (3WC-OR)

Input: Dataset, Number of clusters K, A threshold determination algorithm TDA

Output: Overlapping Region

1: Perform hard clustering on the data to obtain disjoint clusters $c_1, c_2, ..., c_K$

2: **for** $i = 1$ **to** K

3: Use a TDA for determining thresholds

4: 3WC for cluster c_i based on determined thresholds

5: **end**

6: Overlapping Region $= U - \bigcup_{i=1}^{K} Inside(c_i)$

7: Return Overlapping Region

The algorithm needs a dataset, the number of clusters K as inputs and a thresholds determination algorithm. We present two thresholds determination algorithms in the next subsection based on genetic algorithms and game-theoretic rough sets that can be incorporated in the proposed 3WC-OR algorithm. The first step of the algorithm is to apply a conventional hard clustering algorithm such as k-means on the given dataset. Step 2 initializes the iterative process that continues till Step 5. In Step 3, a threshold determination algorithm (TDA) is used to determine thresholds for a certain cluster. We discuss two algorithms in this regards in the next section based on genetic algorithms and game-theoretic rough sets. Each of these algorithms employs an iterative learning process to determine the thresholds. In Step 4, the determined thresholds are used to perform 3WC. Once the 3WC for all the clusters are determined, the next step is to compute the overlapping region by subtracting the universal set from the union of the inside region of all clusters. The output of the algorithm is the overlapping region.

4.1. Genetic Algorithms for Determining 3WC Thresholds

Genetic Algorithms are techniques inspired from natural evolution processes such as selection, crossover and mutation for searching an optimal solution. The genetic algorithm has been used for determining thresholds of three-way decisions based on probabilistic rough set in [23]. In this study, we use it for determination of thresholds defining 3WC.

Algorithm 2 is being constructed based on generalized genetic algorithm called Holland's genetic algorithm [16], for determining thresholds. The algorithm randomly initializes the two thresholds from their respective domains, i.e., D_α and D_β and σ is a variance based criteria that we want to optimize. The inputs of $crossover\_rate$ and $mutation\_rate$ are used inside the algorithm which we explain in the next paragraph.

In Step 1, the algorithm initializes the population which comprises of chromosomes. Each chromosome is a binary encoded representation of a possible (α, β) threshold values. For instance, a three bit binary encoded representa-

Algorithm 2. Genetic Algorithm based Thresholds Determination Algorithm

Input: D_α, D_β, *crossover_rate*, *mutation_rate*, *max_iteration*, σ

Output: Determined thresholds

1: Initialize population by encoding threshold pairs using binary bit strings
2: Evaluate population using σ
3: **Repeat**
4: $(\alpha', \beta') = (\alpha, \beta)$
5: Select next population using evaluated values
6: Perform crossover based on *crossover_rate*
7: Perform mutation based on *mutation_rate*
8: Let $\sigma(\alpha, \beta)$ be the chromosomes with best evaluation from existing population
9: **Until** *max_iteration* > iterations && $\sigma(\alpha, \beta) > \sigma(\alpha', \beta')$
10: Return thresholds (α, β)

tion of threshold α may be "111" = 1, "110" = 0.95, "101" = 0.90 and for β it may be "111" = 0.35, "110" = 0.30 and so on. Each chromosome is a binary string that corresponds to a certain thresholds, for instance the chromosome "111111" will corresponds to thresholds of $(\alpha, \beta) = (1, 0.35)$. Step 2 initializes the iterative process which continues till Step 8. In Step 3, all the chromosomes in the population are evaluated based on the considered criterion σ. In Step 4, new population of chromosomes is generated based on the chromosomes in previous population having higher evaluation values. The selection mechanisms such as roulette-wheel may be used for this purpose. In Step 5, the operation of crossover is carried out. This operation depicts the exchange of genetic information between two parents that lead to an offspring having properties of both parents. The crossover rate is controlled by the input parameter of *crossover_rate*. In Step 6, the mutation operation is carried out which alters one or more gene values in a chromosome from its initial state. The input parameter of *mutation_rate* is used in this step to control that rate of mutation. Step 7 selects the best chromosome from the entire population and defines it

as the (α, β) threshold values. In Step 8, the algorithm checks for the termination of the iterative process. We consider two termination condition i.e., when some specified number of iterations are reached or when there is no improvement in the optimization criterion. In summary, the algorithm will select new chromosomes and therefore new population based on the previously evaluated best chromosomes from the previous population.

4.2. Game-theoretic Rough Sets for Determining 3WC Thresholds

The game-theoretic rough sets provide a game-theoretic formulation for determining a tradeoff solution between multiple criteria that are realized as game players [15, 37]. The players participate in the game by considering strategies in the form of changes in thresholds to improve the overall quality of three-way decisions.

Algorithm 3. Game-theoretic Rough Sets Based Thresholds Determination Algorithm

Input: D_α, D_β, *max_iteration*, σ_1, σ_2

Output: Determined thresholds

1: Randomly initialize α and β such that $\alpha \in D_\alpha$, $\beta \in D_\beta$
2: **Repeat**
3: $(\alpha', \beta') = (\alpha, \beta)$
4: Compute game strategies for the current iteration
5: Computer players σ_1 and σ_2 payoffs based on respective strategies
6: Populate the payoff table and perform equilibrium analysis
7: Determine thresholds (α, β) based on equilibrium
8: **Until** *max_iteration* $>$ iterations && $\sigma(\alpha', \beta') > \sigma(\alpha, \beta)$
9: Return thresholds (α, β)

Algorithm 3 presents a game-theoretic rough sets based algorithm for determining the thresholds. The inputs of the algorithm are same as in genetic algorithm. The algorithm however requires two criteria instead of one which are the two players in the game. In Step 1, the algorithm initializes the two

thresholds from their respective domains. Step 2 initiates the iterative process which continues till Step 6. In Step 3, the thresholds are iteratively updated. In Step 4, game strategies are computed for the current iteration. The game strategies reflect the changes in thresholds and are computed based on improvement in criteria from the previous iteration. We discuss this in detail in our earlier paper [1]. In Step 5, σ_1 and σ_2 payoffs are computed based on their respective strategies. In Step 6, the payoff table is populated and equilibrium analysis is performed on the populated table. In Step 7, thresholds (α, β) are determined based on the determined equilibrium in Step 6. In Step 8 we have similar stop conditions as that of Algorithm 2.

The game-theoretic rough sets based algorithm learns iteratively and in each iteration it selects the best thresholds from the payoff table. The game-theoretic rough sets provides a tradeoff solution while the genetic algorithm provides optimization solution. Moreover, the game-theoretic rough sets consider multiple criteria while the genetic algorithm considers single criterion for determining the thresholds.

5. Experimental Results and Discussion

In this section, we present detailed experimental results of the proposed 3WC-OR algorithm incorporating the variance based criteria. Please note that based on the threshold determination approaches, we consider two implementation of the 3WC-OR algorithm. The implementation of 3WC-OR that considers genetic algorithm for determining thresholds will be referred to as 3WC-OR_{GA} and the implementation of 3WC-OR that considers the game-theoretic rough sets based algorithm for determining the thresholds as 3WC-OR_{GTRS}. The results of these two implementations of 3WC-OR are evaluated on two types of datasets from the UCI machine learning repository namely, single labeled datasets and multi-labeled datasets. The single labeled datasets are those datasets where an instance belongs to exactly one class and therefore the underlying clusters do not contain overlapping instances or regions. We included these

datasets in the experiments to have some intuition regarding the results of the proposed algorithm in situations when there is no overlap among the clusters. The second type of datasets which we refer to as multi-label datasets contain some instances that belong to multiple classes and therefore the underlying clustering results will have overlapping regions. We aim for acceptable results of the proposed algorithm for both the single and multi-labeled datasets [18]. For the single labeled datasets, we use Iris, Wisconsin Breast Cancer and Wine datasets which will be referred to as SD1, SD2 and SD3, respectively. For multi-labeled datasets we considered the Scenes and Birds datasets which will be referred to as MD1, MD2, respectively.

For assessing the quality of the clusters, we use four evaluation measures that are commonly used for evaluation of clusters. The first of these measures is *Accuracy* which is defined as,

$$Accuracy = \frac{\text{Correctly clustered objects}}{\text{Total clustered objects}}. \tag{42}$$

Accuracy means how much accurately we cluster the objects. For measuring the accuracy, we consider one to one matching between clusters and classes. An object is correctly clustered if it belongs to a cluster matching its respective class. The second measure we consider is the *Generality* which is defined as,

$$Generality = \frac{\text{Clustered objects}}{\text{Total objects}}. \tag{43}$$

Generality refers to percentage of objects that are being clustered. An object is not clustered if it not in the inside region of any cluster. The next measure is Davies Bouldin Index which is given by,

$$DB = \frac{1}{K}\sum_{i=1}^{K} \underset{j\neq i}{max}\Big(D_{i,j}\Big), \tag{44}$$

where K is the total number of clusters and $D_{i,j}$ is define as,

$$D_{i,j} = \frac{(\bar{d}_i + \bar{d}_j)}{d_{i,j}}. \tag{45}$$

where $\bar{d}_i$ is the average distance between each point in the i_{th} cluster and the centroid of the i_{th} cluster and $\bar{d}_j$ is the average distance between each point in

the j_{th} cluster and the centroid of the j_{th} cluster. Moreover, $d_{i,j}$ is the distance between the centroids of the i_{th} and j_{th} clusters. Clustering results with least DB value is considered better. The last measure we consider is Silhouette which is given by,

$$s(o_i) = \frac{b(o_i) - a(o_i)}{max\{a(o_i), b(o_i)\}}, \tag{46}$$

where $a(o_i)$ is the average distance of object o_i to all other objects of its own cluster. The $b(o_i)$ is defined by considering the average distance of object o_i from all other objects in a different cluster of which o_i is not a member. The cluster with the minimum average distance is considered as $b(o_i)$. Silhouette provides a representation of how close or similar an object is to its own cluster compared other clusters [31]. Silhouette ranges from -1 to +1. A high value indicates that the object is well matched to its own cluster and poorly matched to other clusters.

In addition to these measures, we also use additional three measures which specifically evaluates the three-way clustering [29]. The first measure is denoted as $Correct_{(inside)}$ and is defined as,

$$Correct_{(Inside)} = \text{Correctly clustered objects by the inside regions.} \tag{47}$$

We consider one to one matching between clusters and classes. An object is correctly clustered in the inside region if it is in the inside of a cluster matching its respective class.

$$Correct_{(Inside,Partial)} = \text{Correctly clustered objects by the inside and partial regions} \tag{48}$$

An object is correctly clustered in the partial region if it is in the partial region of a cluster matching its respective class.

$$Incorrect_{(Inside)} = \text{Incorrectly clustered objects} \tag{49}$$

An object is incorrectly clustered if it is in the inside of a cluster not matching its respective class. Please note that different names were being given to these measures in [29].

5.1. Experimental Results for Single-label Clustering

In this section, we present detailed evaluation results of the single labeled datasets. Before discussing the results, it will be useful to have some comments on the initial configuration and behavior of the two threshold learning algorithms.

The game-theoretic rough set algorithm starts with an initial setting of thresholds $(\alpha, \beta) = (1, 0)$. This selection of thresholds is based on the experimental results and analysis from the previous literature [36]. The genetic algorithm starts with random values for the (α, β) thresholds and considers a fixed chromosome length of 20 bits. Moreover, the optimization criterion in this algorithm is set as the ratio of the within region variance and between region variance mention in Equation (31). In case of game-thoeretic rough sets algorithm, the criteria of within region and between region variance are combined in a game by considering the two criteria as game players. The strategies and the respective threshold modifications are kept the same as in our previous work [1].

Table 3 summarizes the experimental results for the three single labelled datasets. For the sake of comparisons, we also include the results for extreme thresholds settings of (1, 0) and (0.5, 0.5) and we will refer to them as (1, 0) and (0.5, 0.5) models. The columns of the tables show the results of different evaluation measures introduced earlier in the same section. We also included an additional evaluation measure that shows the value of the optimization criterion (introduced in section 3) based on the thresholds used in each of the two implementations of 3WC-OR algorithm.

We will first explain the results of the cluster evaluation measures. We may note that the (1, 0) and (0.5, 0.5) does not necessarily produce the optimized results for the considered criterion. The 3WC-OR$_{\text{GA}}$ provides the best values for the optimization criterion. All the approaches have very high accuracy except the (0.5, 0.5) model. We may also note that the implementations of 3WC-OR produce better generality compared to the (1, 0) model. Across the three datasets, the least generality for 3WC-OR corresponds to 3WC-OR$_{\text{GA}}$ with which is still higher than the (1, 0) model across the three datasets by

a values of 9%, 3% an 9% for SD1, SD2 and SD3, respectively. Although the (0.5, 0.5) model produces optimal generality, however, its accuracy is lesser than other approaches by approximately 10%, 4% and 5% for the three datasets. Moreover, it has the least Silhouette values compared to all other approaches. It is important to note that the single labeled datasets we considered contains objects which can be nicely arranged in clusters and contains very few noisy instances. This can be further confirmed based on accuracy values for the approaches. As a result, when we cluster more objects, the ratio of between to within region variance increases which leads to high values for the DB index. Since with the (0.5, 0.5) model clusters all the objects, therefore in these datasets it leads to a higher value of DB index. The (1, 0) model produces the best results for the Silhouette measure. This is because the (1, 0) model is very restrictive for inclusion of objects in clusters which leads to a high value of Silhouette measure. The 3WC-OR implementations always produce a Silhouette value that is between the Silhouette value obtained with (1, 0) and (0.5, 0.5) models.

Table 3: Results for SD1

		Cluster Evaluation Measures				3WC Evaluation Measures			
		$\frac{\sigma_B^2}{\sigma_W^2}$	Accuracy	Generality	DB	Silhouette	$Co_{\cdot(In.)}$	$Co_{\cdot(In.,Pa.)}$	$Inc.$
SD1	3WC-OR$_{\text{GTRS}}$	5.2136 ± 0.862	0.9864 ± 0.0191	0.9221 ± 0.0231	0.5947 ± 0.006	0.7825 ± 0.006	138	**143**	4
	GA-TWC$_{\text{GA}}$	**6.8261** ± 0.6201	0.9816 ± 0.0308	0.8427 ± 0.0373	0.5088 ± 0.0097	0.8255 ± 0.0097	138	138	9
	(1, 0) model	5.247 ± 0.454	**0.9895** ± 0.0215	0.7372 ± 0.0361	**0.4152** ± 0.0041	**0.8731** ± 0.0089	110	130	**0**
	(0.5, 0.5) model	6.0805 ± 0.625	0.8871 ± 0.0326	**1** ± 0	0.657 ± 0.0082	0.7297 ± 0.006	**140**	140	10
SD2	3WC-OR$_{\text{GTRS}}$	0.8843 ± 0.6037	0.9678 ± 0.0115	0.9056 ± 0.0139	0.7034 ± 0.0036	0.8035 ± 0.0036	**656**	656	20
	3WC-OR$_{\text{GA}}$	**0.9791** ± 0.672	**0.9925** ± 0.0185	0.8304 ± 0.0224	0.6329 ± 0.0058	0.8585 ± 0.0058	**656**	**657**	22
	(1, 0) model	0.892 ± 0.526	0.9915 ± 0.0131	0.7974 ± 0.0192	**0.6321** ± 0.045	**0.8612** ± 0.0048	564	622	**4**
	(0.5, 0.5) model	0.8341 ± 0.6213	0.93982 ± 0.0235	**1** ± 0	0.8109 ± 0.0032	0.7239 ± 0.0043	650	650	33
SD3	3WC-OR$_{\text{GTRS}}$	0.8707 ± 0.3744	0.963 ± 0.0091	0.7927 ± 0.0107	0.9675 ± 0.0022	0.5918 ± 0.0019	146	150	12
	3WC-OR$_{\text{GA}}$	**1.3336** ± 0.5418	**0.9946** ± 0.0146	0.5716 ± 0.0172	**0.7048** ± 0.0035	0.7247 ± 0.0031	154	154	15
	(1, 0) model	1.0089 ± 0.2371	0.9908 ± 0.0139	0.4863 ± 0.0143	0.7098 ± 0.0029	**0.7250** ± 0.0026	87	132	**0**
	(0.5, 0.5) model	1.062 ± 0.3216	0.9172 ± 0.0025	**1** ± 0	1.195 ± 0.0121	0.4825 ± 0.0015	**160**	**160**	18

We now explain the results of the 3WC evaluation measures. For all the three datasets, we may note that the (0.5,0.5) model tends to have more number of correct objects in the inside regions and have zero objects in the partial region, however, the same model also has the highest number of incorrect ob-

ject placements. The 3WC-OR approaches tend to have slightly lesser correct objects in the inside region with a slight increase in boundary but have lesser number of incorrect object placements compared to (0.5,0.5) model. When the partial region is also included, (see column $Co._(In., Pa.)$) the number of correct object placements by the 3WC-OR approaches is almost the same as that of the (0.5,0.5) model. The (1,0) model has far lesser correct objects in the inside region and has many objects in the partial region. It however has least number of incorrect object placements. In conclusion, the 3WC-OR approaches may be considered as some sort of tradeoff between the two extreme settings of thresholds. It avoids many objects in the partial region while having lesser number of incorrect assignments.

Table 4: Accuracy comparison of different algorithms for SD1, SD2 and SD3

Algorithm	SD1	SD2	SD3	Algorithm	SD1	SD2	SD3
HCM [22]	0.8187	0.9219	0.9407	RECM [24]	0.8506	0.9286	0.9343
SC [26]	0.8555	0.9262	0.9663	CE3 [33]	0.9067	0.9501	0.9831
FCM [6]	0.8453	0.9279	0.9494	3WC-OR$_{\text{GTRS}}$	0.9485	0.9236	0.8670
RFCM [25]	0.7983	0.9306	0.9561	3WC-OR$_{\text{GA}}$	0.9058	**0.8985**	**0.7809**

To better assess the proposed 3WC-OR algorithm, we report the comparison of the proposed approaches with some of the existing clustering methods on the same datasets. An important issue while evaluating the performance of the 3WC algorithms is that it tends to produce higher accuracy. This is because some hard to cluster points (whose association with a cluster are not clear) are not clustered and are assigned to partial region. This results in high accuracy results thereby making it difficult to make comparison of 3WC approaches with other existing clustering methods. To have a better intuition of the relative performance of the 3WC approaches with existing methods, we consider a measure that is reported in our earlier work [36]. The measure computes the accuracy by considering a worst case scenario where we will make completely random decisions about the objects in the partial region. It is further assumed that for all the objects in the partial region, we have a 50% chance of making a correct cluster assignment

and an equal 50% chance of making an incorrect cluster assignment. The new accuracy is computed as $(Accuracy \times Generality) + (0.5 \times 1 - Generality)$. This formula computes the weighted mean of the accuracy in the clustered and non clustered regions. In the clustered region, the accuracy is based on the values of Table 3 while in the non clustered region the accuracy is assumed to be 50%. We report this accuracy for the sake of comparison.

Table 4 presents the comparison results. We may note that 3WC-OR$_{\text{GTRS}}$ provides better results on SD1 compared to the to the approaches. In case of SD2 the results are comparable to most of the approaches. In case of SD3, however the results are a bit on the lower side. This is however acceptable since we have considered accuracy in the worst case scenario. Moreover, with regards to the CE3 which are is also a 3WC (approach having better accuracy in case of SD2 and SD3), the reported accuracy results are based on the clustered objects only. It is also of interest to note that on the SD1 and SD3 the results of the measures $Correct_{(Inside)}$, $Correct_{(Inside,Partial)}$ and $Incorrect_{(Inside)}$ are also reported in the work of Peters in [29]. The best result for $Correct_{(Inside)}$ in their case on SD1 and SD3 are 122 and 136 correct objects, respectively. The values of $Incorrect_{(Inside)}$ objects being in the range of 1 to 3 for the two dataset. Comparing this to our approach, the best values of $Correct_{(Inside)}$ in our case are 138 and 154 correct objects for SD1 and SD2, respectively. The number of incorrect objects are however higher in our case. This is due to the fact that the approaches introduced by Peters in [29] tend to avoid clustering of many objects therefore leading to larger partial region while our approach tries to cluster more objects and looks for reducing partial region.

5.2. Experimental Results for Multi-label Clustering

To get more better intuition of the performance and evaluation of the proposed 3WC-OR approaches, we introduce additional measures that are specifically designed for evaluation of multi-label datasets. Considering, TL_i and PL_i be the sets of true labels and predicted labels for an object o_i, the $F1$ measure

for multi-label datasets having N objects is defined as,

$$F1 = \frac{1}{N}\sum_{i=1}^{N}\frac{2\times|TL_i\bigcap PL_i|}{|TL_i|+|PL_i|} \tag{50}$$

Moreover, we also include the measure of hamming loss which is defined as,

$$HammingLoss = \frac{1}{KN}\sum_{i=1}^{N}|TL_i\bigcap PL_i^c|+|TL_i^c\bigcap PL_i| \tag{51}$$

where K is the number of classes or equivalently the number of clusters and PL_i and TL_i are the set complements of the respective sets. The hamming loss measures the prediction error and also the missing error, i.e., the categories which are not predicted.

Table 5 presents the experimental results for the multi-labeled datasets namely Scenes and Birds which we refer to as MD1 and MD2, respectively. In case of MD1, the (0.5, 0.5) model leads to better optimization of the proposed criterion. The implementations 3WC-OR are unable to determine such thresholds because the stopping conditions in these implementations do not allow for such configuration of thresholds, i.e., $(\alpha, \beta) = (0.5, 0.5)$ and works strictly within the majority oriented setting of thresholds that is $1.0 \leq \alpha < \beta \leq 0.0$. However, in case of MD2, the 3WC-OR$_{\text{GA}}$ provides better results for the optimization criterion. In both datasets, the implementation of 3WC-OR have similar accuracy

Table 5: Results of MD1 and MD2

Dataset	Algorithms	Typical evaluation measures					Multi-label Measures	
		$\frac{\sigma_B^2}{\sigma_W^2}$	Accuracy	Generality	DB	Silhouette	F1-Measure	Hamm.Loss
MD1	3WC-OR$_{\text{GTRS}}$	0.2315 ± 0.0915	0.8835 ± 0.0192	0.8811 ± 0.021	2.1015 ± 0.0055	0.1797 ± 0.0183	0.7011 ± 0.0147	0.1185 ± 0.0192
	3WC-OR$_{\text{GA}}$	0.3423 ± 0.0473	**0.8943** ± 0.0221	0.7707 ± 0.0339	2.493 ± 0.0088	**0.2395** ± 0.0295	**0.7424** ± 0.0223	**0.0987** ± 0.0221
	(1, 0) model	0.0853 ± 0.0021	0.6730 ± 0.0216	0.0245 ± 0.0315	2.2130 ± 0.0042	0.201 ± 0.0512	0.7023 ±0.0147	0.2926 ± 0.0321
	(0.5, 0.5) model	**0.3442** ± 0.0121	0.8941 ± 0.0262	**1** ± 0	**2.9123** ± 0.0036	0.1155 ± 0.0391	0.6949 ± 0.0012	0.1696 ± 0.0153
MD2	3WC-OR$_{\text{GTRS}}$	0.4621 ± 0.0107	0.9408 ± 0.0232	0.7959 ± 0.0455	0.3089 ± 0.0066	**0.9013** ± 0.0221	0.7401 ± 0.0143	0.076 ± 0.0232
	3WC-OR$_{\text{GA}}$	**0.5663** ± 0.0782	**0.9435** ± 0.0267	0.6807 ± 0.071	0.311 ± 0.0107	0.755 ± 0.0356	**0.7825** ± 0.023	**0.0771** ± 0.0267
	(1, 0) model	0.5262 ± 0.0145	0.8903 ± 0.0342	0.1403 ± 0.043	0.1746 ± 0.0025	0.841 ± 0.0158	0.7402± 0.0254	0.1035 ± 0.0236
	(0.5, 0.5) model	0.4785 ± 0.0236	0.9013 ± 0.0294	**1** ± 0	**0.5356** ± 0.0026	0.7594 ± 0.0254	0.7148 ± 0.0191	0.0942 ± 0.0254

results. The accuracy of (0.5, 0.5) model is comparable to that of the 3WC-OR in case of MD1, however, it has 4% lesser accuracy in case of MD2. For the (1, 0) model, the accuracy is significantly low for MD1 while in case of MD2 its

accuracy is 5% lesser than the two implementation of 3WC-OR. The generality for the (1, 0) model is very poor for both the datasets. Again as expected, the generality of the (0.5, 0.5) model has a maximum value for both the datasets. For both the datasets, the Silhouette measure values are not maximum for the (1, 0) model. The 3WC-OR provides better results for the Silhouette measure in both the datasets. The DB index is however, higher for the (0.5, 0.5) model in both the cases. For the multi-labelled measures, the 3WC-OR$_{\mathrm{GA}}$ outperforms the other approaches on both the measures of F1 and hamming loss. The (1,0) model although provides similar F1 results to that of 3WC-OR$_{\mathrm{GTRS}}$, however, in comparison to 3WC-OR$_{\mathrm{GTRS}}$, it has 16% and 3% higher hamming loss on MD1 and MD2 (i.e., Scenes and Birds datasets), respectively. The hamming loss for (1,0) and (0.5,0.5) are always higher than the 3WC-OR approaches.

5.3. Analysis of the Overlapping Region

In this section, we go a bit deeper and look at results of the 3WC-OR algorithm specifically in the overlapping region. In particular, we look at how well the two implementations of 3WC-OR performed in estimating or computing the overlapping region. To perform such analysis, we introduce the following measures,

$$Precision_{OL} = \frac{\text{Correctly ident. objects from overlapping region}}{\text{Total identified objects in overlapping region}}, \quad (52)$$

$$Recall_{OL} = \frac{\text{Correctly ident. objects from overlapping region}}{\text{Total objects in overlapping region}}, \quad (53)$$

$$Accuracy_{OL} = \frac{\text{Correctly ident. overlap. \& non-overlap. objects}}{\text{Total objects}}. \quad (54)$$

The $Precision_{OL}$ reflects that out of the total assigned objects to the overlapping region by a certain approach, how many have been correctly assigned. On the other hand, the $Recall_{OL}$ reflects that out of the total objects that belong to the overlapping region, how many have been correctly assigned or identified. Lastly, the $Accuracy_{OL}$ reflect the percentage correctness of accurately assigning objects to the overlapping and non-overlapping regions.

Table 6: Quality attributes of overlapping region

	$Precision_{OL}$		$Recall_{OL}$		$Accuracy_{OL}$	
	MD1	MD2	MD1	MD2	MD1	MD2
3WC-OR_{GTRS}	**0.4016**	**0.5385**	0.5763	0.5556	0.9057	**0.7302**
3WC-OR_{GA}	0.1875	0.3273	0.0339	0.0952	0.9181	0.6775
(1, 0) model	0.0825	0.297	**1**	**0.9362**	0.182	0.3721
(0.5, 0.5) model	0	0	0	0	**0.9265**	0.707

Table 6 presents values of these measures for MD1 and MD2. The results indicate that among the implemenations, the 3WC-OR_{GTRS} achieved the highest $Precision_{OL}$ and $Recall_{OL}$ for both MD1 and MD2. The 3WC-OR_{GTRS} also produced the best $Accuracy_{OL}$ value for MD2 and slightly lower value than the best $Accuracy_{OL}$ achieved for MD1. The 3WC-OR_{GA} outperforms other approaches in finding the best within to between region ratio, however, the $Precision_{OL}$ and $Recall_{OL}$ values for MD1 and MD2 are not that much effective. From Table 6, we may note that the 3WC-OR_{GA} got the high $Accuracy_{OL}$ of the overlapping region for MD1, however, results for MD2 are a bit on the lower side. $Precision_{OL}$ for (1, 0) model was only 8.25% for MD1 and a slightly better value of 29.7% for MD2. This is due to the fact that the (1, 0) model leads to larger size of the overlapping region (as highlighted in Section 2.1) which leads to lots of incorrect assignment to the overlapping region. On the other hand, it has a very high $Recall_{OL}$ of 100% and 93% for the MD1 and MD2 respectively. This is because the (1, 0) model puts a lot of objects in the overlapping region and therefore the actual overlapping region is more or less the subset of the predicted overlapping region, thereby resulting in a higher $Recall_{OL}$ value. In summary, the (1, 0) model identifies majority of the overlapping objects, but its prediction $Accuracy_{OL}$ or $Precision_{OL}$ for identifying an object as belonging to the overlapping region is not very effective.

The (0.5, 0.5) model has minimum $Precision_{OL}$ and $Recall_{OL}$ values because no objects are being placed in the overlapping region. The $Accuracy_{OL}$

for this model is however, acceptable and somewhat comparable to the other approaches. This is because this model assigns all the objects to the non-overlapping region. Since the considered datasets has a dominant non-overlapping region (single labeled instances), i.e., 92.65% for MD1 or Scenes dataset and 70.7% for MD2 or Birds dataset, therefore the (0.5,0.5) results in high value for $Accuracy_{OL}$. It should be noted that $Accuracy_{OL}$ value of (0.5, 0.5) model is dependent on the number of objects in the overlapping region. If the number of objects in the overlapping region increase the $Accuracy_{OL}$ of (0.5, 0.5) model will decrease and vice versa.

The detailed experimental results presented in this section suggest that the proposed 3WC-OR algorithm can be quite useful for estimating and handling overlapping regions in clustering.

6. Conclusion

In many real applications clusters may not necessarily have clear and sharp boundaries and therefore has an overlapping region. The three-way clustering or 3WC is an effective and useful approach for handling overlapping clustering. The determination of thresholds in 3WC plays a crucial role in accurate estimation of the overlapping region of clusters. In this article, we consider different variance based criteria for determining the thresholds. A three-way clustering algorithm for overlapping region or 3WC−OR, is introduced which considers the determination of thresholds based on approaches such as genetic algorithms and game-theoretic rough sets for effective estimation of the clusters overlapping region. More specifically, the 3WC−OR considers the optimization of variance or spread in evaluation function values of objects from two perspectives, i.e., within region variance and between region variance. The underlying conjecture in the proposed variance based criteria is that well thresholded regions are those that are well separated in evaluation function values. Experimental results and comparisons with some of the existing three-way and other similar approaches suggest that the proposed algorithm provides better estimation of the overlap-

ping region. It provides an increase in accuracy of up to 4% for datasets with overlapping clusters while having similar accuracy results for datasets with non-overlapping clusters.

Acknowledgment

This work was partially supported by Higher Education Commission of Pakistan under the grant Indigenous Ph.D. Fellowship Program and an NSERC discovery grant Canada.

References

[1] Afridi, M.K., Azam, N., Yao, J.T., Alanazi, E., 2018. A three-way clustering approach for handling missing data using GTRS. International Journal of Approximate Reasoning 98, 11–24.

[2] Arabie, P., Carroll, J.D., DeSarbo, W., Wind, J., 1981. Overlapping clustering: A new method for product positioning. Journal of Marketing Research 18, 310–317.

[3] Aslam, J.A., Pelekhov, E., Rus, D., 2004. The star clustering algorithm for static and dynamic information organization. Journal of Graph Algorithms and Applications 8, 95–129.

[4] Azam, N., Yao, J.T., 2016. Variance based determination of three-way decisions using probabilistic rough sets, in: Proceedings of International Joint Conference on Rough Sets (IJCRS'16). Lecture Notes in Computer Science 9920, pp. 209–218.

[5] Bezdek, J.C., 1981. Pattern Recognition with Fuzzy Objective Function Algorithms. Kluwer Academic Publishers, Norwell, MA, USA.

[6] Bezdek, J.C., Ehrlich, R., Full, W., 1984. Fcm: The fuzzy c-means clustering algorithm. Computers & Geosciences 10, 191–203.

[7] Campagner, A., Ciucci, D., 2019. Orthopartitions and soft clustering: Soft mutual information measures for clustering validation. Knowledge-Based Systems 180, 51–61.

[8] Chaturvedi, A., Carroll, J.D., Green, P.E., Rotondo, J.A., 1997. A feature-based approach to market segmentation via overlapping k-centroids clustering. Journal of Marketing Research 34, 370–377.

[9] Cleuziou, G., 2008. An extended version of the k-means method for overlapping clustering, in: 19th International Conference on Pattern Recognition, pp. 1–4.

[10] Deng, X.F., Yao, Y.Y., 2014. A multifaceted analysis of probabilistic three-way decisions. Fundamenta Informaticae 132, 291–313.

[11] Gabrielli, L., Giuffrida, S., Trovato, M.R., 2017. Gaps and overlaps of urban housing sub-market: hard clustering and fuzzy clustering approaches, in: Appraisal: From Theory to Practice, pp. 203–219.

[12] Gao, C., Yao, Y.Y., 2016. Determining thresholds in three-way decisions with chi-square statistic, in: International Joint Conference on Rough Sets (IJCRS'16), Lecture Notes in Computer Science 9920, pp. 272–281.

[13] Gil-García, R.J., Badía-Contelles, J.M., Pons-Porrata, A., 2003. Extended star clustering algorithm, in: Iberoamerican Congress on Pattern Recognition, pp. 480–487.

[14] Goldberg, M., Kelley, S., Ismail, M.M., Mertsalov, K., Wallace, A., 2010. Finding overlapping communities in social networks, in: ieee 2nd international conference on Social computing (socialcom), pp. 104–113.

[15] Herbert, J.P., Yao, J.T., 2011. Game-theoretic rough sets. Fundamenta Informaticae 108, 267–286.

[16] Holland, J.H., 1992. Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. MIT press.

[17] Krishnapuram, R., Joshi, A., Yi, L., 1999. A fuzzy relative of the k-medoids algorithm with application to web document and snippet clustering, in: IEEE Interanationl Conference on Fuzzy Systems, pp. 1281–1286.

[18] Lichman, M., 2013. UCI machine learning repository. URL: `http://archive.ics.uci.edu/ml.(Retrieved:2019-04-06)`.

[19] Lingras, P., Peters, G., 2011. Rough clustering. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 1, 64–72.

[20] Lingras, P., West, C., 2004. Interval set clustering of web users with rough k-means. Journal of Intelligent Information Systems 23, 5–16.

[21] Long, A.N., Dagogo-Jack, S., 2011. Comorbidities of diabetes and hypertension: mechanisms and approach to target organ protection. The Journal of Clinical Hypertension 13, 244–251.

[22] MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations, in: 5th Berkeley symposium on mathematical statistics and probability, pp. 281–297.

[23] Majeed, B., Azam, N., Yao, J.T., 2014. Thresholds determination for probabilistic rough sets with genetic algorithms, in: International Conference on Rough Sets and Knowledge Technology (RSKT'14), Lecture Notes in Computer Science 8818, pp. 693–704.

[24] Masson, M.H., Denœux, T., 2009. Recm: Relational evidential c-means algorithm. Pattern Recognition Letters 30, 1015–1026.

[25] Mitra, S., Banka, H., Pedrycz, W., 2006. Rough–fuzzy collaborative clustering. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 36, 795–805.

[26] Ng, A.Y., Jordan, M.I., Weiss, Y., 2002. On spectral clustering: Analysis and an algorithm, in: Advances in neural information processing systems, pp. 849–856.

[27] Pedrycz, W., 2005. Interpretation of clusters in the framework of shadowed sets. Pattern Recognition Letters 26, 2439–2449.

[28] Pérez-Suárez, A., Martínez-Trinidad, J.F., Carrasco-Ochoa, J.A., Medina-Pagola, J.E., 2013. Oclustr: A new graph-based algorithm for overlapping clustering. Neurocomputing 121, 234–247.

[29] Peters, G., 2014. Rough clustering utilizing the principle of indifference. Information Sciences 277, 358–374.

[30] Peters, G., Crespo, F., Lingras, P., Weber, R., 2013. Soft clustering–fuzzy and rough approaches and their extensions and derivatives. International Journal of Approximate Reasoning 54, 307–322.

[31] Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20, 53–65.

[32] Sun, P.G., Gao, L., Han, S.S., 2011. Identification of overlapping and non-overlapping community structure by fuzzy clustering in complex networks. Information Sciences 181, 1060–1071.

[33] Wang, P.X., Yao, Y.Y., 2018. Ce3: A three-way clustering method based on mathematical morphology. Knowledge-Based Systems 155, 54–65.

[34] Whang, J.J., Gleich, D.F., Dhillon, I.S., 2016a. Overlapping community detection using neighborhood-inflated seed expansion. IEEE Transactions on Knowledge and Data Engineering 28, 1272–1284.

[35] Whang, J.J., Gleich, D.F., Dhillon, I.S., 2016b. Overlapping community detection using neighborhood-inflated seed expansion. IEEE Transactions on Knowledge and Data Engineering 28, 1272–1284.

[36] Yao, J.T., Azam, N., 2015. Web-based medical decision support systems for three-way medical decision making with game-theoretic rough sets. IEEE Transactions on Fuzzy Systems 23, 3–15.

[37] Yao, J.T., Herbert, J.P., 2008. A game-theoretic perspective on rough set analysis. Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition) 20, 291–298.

[38] Yao, Y., Lingras, P., Wang, R., Miao, D., 2009. Interval set cluster analysis: a re-formulation, in: International Workshop on Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing, pp. 398–405.

[39] Yao, Y.Y., 2012. An outline of a theory of three-way decisions, in: 8th International Conference on Rough Sets and Current Trends in Computing (RSCTC'12), Lecture Notes in Computer Science 7413, pp. 1–17.

[40] Yao, Y.Y., Gao, C., 2015. Statistical interpretations of three-way decisions, in: 10th International Conference on Rough Sets and Knowledge Technology (RSKT'15), Lecture Notes in Computer Science 9436, pp. 309–320.

[41] Yu, H., 2017. A framework of three-way cluster analysis, in: 2nd International Joint Conference on Rough Sets (IJCRS'17), Lecture Notes in Computer Science 10313, pp. 300–312.

[42] Yu, H., Chen, Y., Lingras, P., Wang, G., 2019. A three-way cluster ensemble approach for large-scale data. International Journal of Approximate Reasoning 115, 32–49.

[43] Yu, H., Jiao, P., Yao, Y.Y., Wang, G.Y., 2016. Detecting and refining overlapping regions in complex networks with three-way decisions. Information Sciences 373, 21–41.

[44] Yu, H., Liu, Z., Wang, G., 2014a. An automatic method to determine the number of clusters using decision-theoretic rough set. International Journal of Approximate Reasoning 55, 101–115.

[45] Yu, H., Su, T., Zeng, X.H., 2014b. A three-way decisions clustering algorithm for incomplete data, in: 9th International Conference on Rough Sets and Knowledge Technology (RSKT'14), Lecture Notes in Computer Science 8818, pp. 765–776.

[46] Zamir, O., Etzioni, O., 1998. Web document clustering: A feasibility demonstration, in: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 46–54.

[47] Zhang, M.L., Zhou, Z.H., 2007. Ml-knn: A lazy learning approach to multi-label learning. Pattern recognition 40, 2038–2048.

[48] Zhang, S., Wang, R.S., Zhang, X.S., 2007. Identification of overlapping community structure in complex networks using fuzzy c-means clustering. Physica A: Statistical Mechanics and its Applications 374, 483–490.

[49] Zhou, J., Lai, Z., Miao, D., Gao, C., Yue, X., 2020. Multigranulation rough-fuzzy clustering based on shadowed sets. Information Sciences 507, 553–573.