Pertanika J. Sci. & Technol. 25 (S): 211 - 222 (2017)

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Local Outlier Factor in Rough K-Means Clustering

Khaled Ali Othman*, Md Nasir Sulaiman, Norwati Mustapha and Nurfadhlina Mohd Sharef

Department of Computer Science, Faculty of Computer Science and IT, University Putra of Malaysia, Selangor, Malaysia

ABSTRACT

K-Means is an unsupervised method partitions the input space into clusters. K-Means algorithm has a weakness of detecting outliers, which have it available in many variations research fields. A decade ago, Rough Sets Theory (RST) has been used to solve the problem of clustering partition. Specifically, Rough K-Means (RKM) is a one of the powerful hybrid algorithm, which has it, has various extension versions. However, with respect of the ideas of existing rough clustering algorithms, a suitable method to detect outliers is much needed now. In this paper, we propose an effective method to detect local outliers in rough clustering. The Local Outlier Factor (LOF) method in rough clustering improves the quality of the cluster partition. The improved algorithm increased the level of clusters quality. An existing algorithm version, the π Rough K-Means (π RKM) tested in the study. Finally, the effectiveness of the algorithm performance is demonstrated based on synthetic and real datasets.

Keywords: Data analysis, Local outlier factor, K-Means; Rough clustering, Outlier detection

INTRODUCTION

Clustering/cluster analysis has a deep and wealthy history of more than 50 years in various scientific fields and widely used until now (Jain, 2010; Berkhin, 2006). Clustering is an unsupervised learning technique concerned for grouping records or samples of data sets. K-Means is a simple and popular partition clustering algorithm, has been successfully used in various application domains, such as information retrieval, image segmentation and among others (Lingras, 2007). However, K-Means found to be weak in detecting outliers, where continuous attempts to find solution has been conducted in various research fields.

ARTICLE INFO

Article history:
Received: 15 August 2016
Accepted: 18 May 2017

E-mail addresses:
khaledgs37650@gmail.com (Khaled Ali Othman),
nasir@upm.edu.my (Md Nasir Sulaiman),
norwati@upm.edu.my (Norwati Mustapha),
Nurfadhlina@upm.edu.my (Nurfadhlina Mohd)
*Corresponding Author

ISSN: 0128-7680 © 2017 Universiti Putra Malaysia Press.

Over the last decade, in an intelligent computing field various approaches (including fuzzy logic, rough sets, neural network, and genetic algorithm) has been evaluated and applied to handle different challenges posed by data analysis. Rough set is a popular approach has been used to solve the problem of cluster partitions quality. For instance, the basic rough set properties applied with K-Means algorithm to separate the clusters clearly (an object either belongs to one or more than one "crisp" cluster). The combination of basic rough properties and the K-Means algorithm yields hybrid algorithm which was introduced by Lingers and West (Lingras & West, 2004). Therefore, Rough K-Means algorithm (RKM) is proposed for characterizing overlapping objects from partitioning clusters. Two approximations introduced for each cluster, which called a lower (also known as positive region) and an upper approximation (also known as negative region) as a solution. All the objects in positive region are belong to one cluster. On the contrary, all objects in the negative regions are possibly belong to two or more clusters (a brief description of each region is explained Section 2). Thus, the rough clustering approach provides a new insight to improve the quality of clustering. In conclusion, rough clustering affirmed as a popular and active approach in various application domains.

In recent years, concern on rough clustering, especially RKM (Peters, 2014) has been emerging. The powerful algorithm has relevantly improvised by applying some refinement as proposed by Peters (Peters, 2006) and further ascertained by Lingers & Peters (Lingras & Peters, 2012) study. The literature by Peters (Peters, 2014) proposed a new method to calculate the means by using Laplace's principle of indifference. However, on the basis of the evidence currently available seems to suggest that a suitable method to detect outliers is needed. The consensus view seem to be that, detecting outliers based on the distance of an object from mean (Centroid) in another cluster may not quite efficient. This explains the fact that each cluster has its own density and the deeper outlier in each cluster. Our contribution is to evaluate RKM clustering algorithm with a different method to detect outliers. In other words, the local outlier factor method may also be adapted as an effective measure to detect outliers in rough clustering. The claim is demonstrated on synthetic and real datasets from Breast Cancer Wisconsin and Iris Plant.

The remainder of the paper is organized into sections: Section II introduces the related work; Section III discusses the conception of identifying the local outlier factor; Section IV discusses the experimental evaluation; and Section V concludes the paper.

MATERIALS AND METHODS

Rough Sets – Basic Concepts

Rough Set Theory (RST) is a mathematical formalism to treat the vagueness and uncertain information (Pawlak, 1982) in soft computing field. RST uses classification to treat uncertain or incomplete information in the data. This idea has been successfully applied in many research areas by classifying a set of objects based on an approximation space.

Rough Approximations

Approximations are a fundamental construct that distinguishes a rough set from the other approaches. The essential idea of approximation is to isolate indiscernible form discernible objects (definition of lower and upper approximation expressed in the introduction above). Figure1 depicts a definition of approximation in the RST.

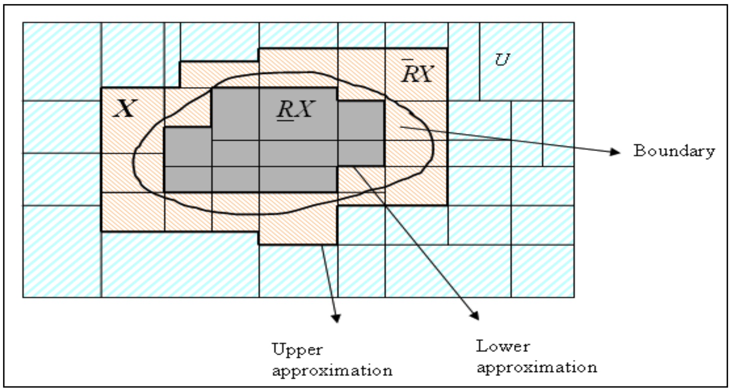

Figure 1. Definitions of approximation in RST

The lower $\underline{R}(X)$, and upper $\overline{R}(X)$ approximation can divide the Universe (U) into the following regions:

- Positive (POS(X)=$\underline{R}(X)$),
- Negative (NEG(X)=$\overline{R}(X)$), and
- Boundary (BND(X)=$\overline{R}(X)$ - $\underline{R}(X)$).

We need to mention here that, there are connections and differences between the concept of rough sets and fuzzy sets. Generally, rough set and fuzzy set are both classical theories for modeling the vague and the imprecise information. However fuzzy set involves more advanced mathematical concepts and used to derive structures, numbers and functions by employing the fuzzy membership function. Indeed rough set has an advantage in data analysis, which it does not need any preliminary or additional information about data (Peters, Crespo, Lingras, & Weber, 2013).

Rough Clustering. In rough clustering approach, basic properties of rough set is required (Peters, 2006). They are as follows:

- An object X which belongs to one lower approximation must not be overlap with another approximation.
- An object X that is a member of a lower approximation is a subset of its corresponding upper approximation.

- If an object X does not belonging to any lower approximation, it belongs to two or more upper approximations.

In reality these “basic properties are not necessarily independent or complete” (Lingras & Peters, 2012). But, it is notable that basic properties do provide an understanding on rough set adaption in clustering partitions techniques such as K-Means algorithm. Figure 2 explains the rough boundaries based on basic properties of rough set using three clusters.

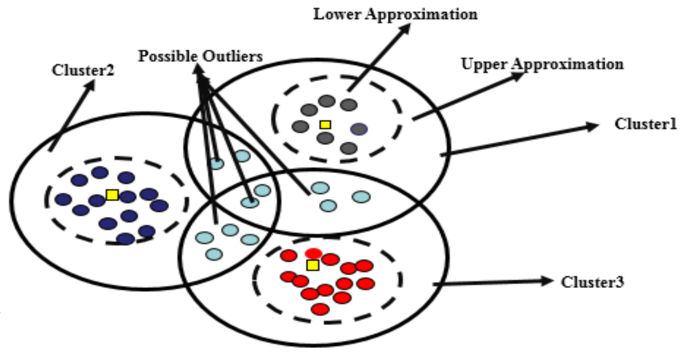

Figure 2. Example Rough Clustering boundary

Rough K-Means (RKM) Versions. Despite various studies on (RKM) the main essential’s effort of RKM clustering algorithm proposed are (Lingras & West, 2004); (a) calculation of the mean (Centroids) and (b) an object assigned to the lower/upper approximations three factors are identified as inputs. First, estimate the number of clusters. However, finding the appreciate number of clusters is more difficult and it’s just based on a trial or error process. Second, two weights corresponded as parameters (W_l, W_u), which are represent the linear combination of lower and upper means (see (Lingras & Peters, 2012) for more details on calculation of the means). The third factor is to determine the size of the boundaries by using a threshold (T). At this point, the numbers of objects in the boundary region would be decreased by increasing the value of threshold.

In addition, Peters (Peters, 2006) has done some refinements in RKM algorithm. He studied and proposed some alternative methods to improvised algorithm. The proposed method is the weight for lower approximations and upper approximations regions, where W_u=1-W_l (generally set W_l=0.7). He too applied relative distance rather than Laplace’s distance method to detect overlapping. The main concept of using relative distance method in assigning the object to the lower or boundary region is described in Figure 3.

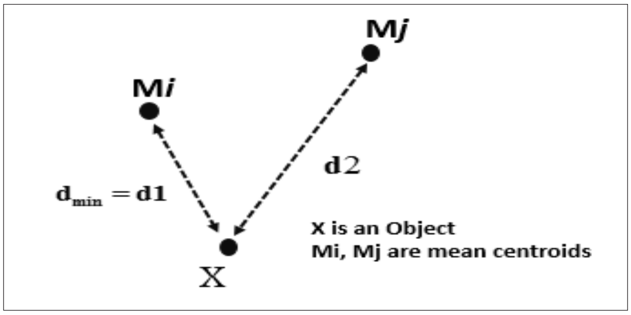

Figure 3. Assigning an object to an approximation

Figure 3 assumes X as an object and M_i and M_j are two mean clusters carotids. The d1 is the minimum distance (d_{min}) between the object X and closest mean (Centroid) M_i. Meanwhile, d2 is a distance between the object X and other means M_j . The equation to determine if the object is overlap or non-overlap is computed as follows:

$$T = \left\{ \left(d2 \big/ d_{\min} \right) \leq T \right\} \qquad [1]$$

Recently, an important refinement of RKM algorithm has been presented by Peter (Peters, 2014). He discussed RKM algorithm and proposed a new mean functional by using Laplace's principle of indifference method. Peter (Peters, 2014) indicates that the numbers of objects in the lower and boundary approximation are neglected. The π Rough K-Means (πRKM) algorithm is an existing algorithm version, which enhanced the results of partition cluster quality. The πRKM algorithms steps involve are follows:

Input: K Numbers, Threshold.

Output: rough Clusters.

Step 0: Initialization.

- Determine the initial means (e.g., randomly or maximum distance between means).
- Assign each object Xn to the corresponding upper approximation of its nearest mean.

Step 1: Compute the new means as follows:

$$M_k = \left(\sum_{xi \in \overline{R}i)}^{n} \frac{Xi}{\overline{R}i} \right) / \left(\sum_{xi \in \overline{R}i)}^{n} \frac{1}{\overline{R}i} \right)$$

Step 2: Assign the objects to the approximations:

- Determine the nearest Centroid:

  d_{min}= d(X,M_i) = min1≤j≤kd(X,M_j).

- Determine if further data object also closest to other centroids or not by using relative distance and threshold as defined below:

  Let T' = {j: $d(X,M_j)/d(X,M_i) \leq$ threshold and $i \neq j$}. Then we get:

  - If T' $\neq \emptyset$ then at least one other centroid is similarly close to the object.
  - If T' $= \emptyset$ then no other centroids are similarly close to the object.

Step 3: Check convergence of the algorithm.

- If the algorithm has not converged continue with Step 1.
- Else STOP.

Discussion on RKM Algorithm outlier detection Method. As mentioned earlier in RKM algorithm, the relative distance method has been used as measure to detect cluster overlaps. The idea of RKM detecting outliers is depends on how the object in the one clusters far from the mean centroid in other cluster. However, RKM does not consider on how the isolate data objects in the cluster affect the quality of the cluster. With reference to the nature of the original data and their density, the quality of clusters partitions is affected by the nature of data density. For instance in Figure 4, C2 is depicts a dense cluster less than C1. The increasing or decreasing of RKM detection outlier threshold may affect the quality of clusters.

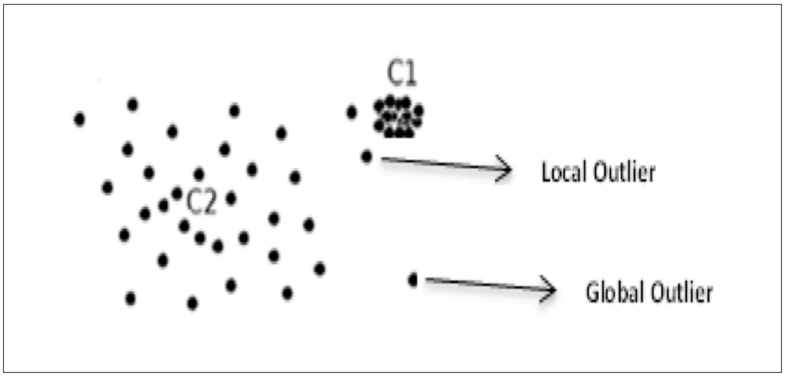

Figure 4. Example of outliers based on two clusters

Local Outlier Factor Definition. Local Outlier Factor (LOF) is a ratio to estimate reachability density of the area around the object to the local =densities of its neighbors. The successful method has been used originally =with a hierarchical clustering algorithm, namely OPTICS (Ordering Points to =Identify the Clustering Structure) (Breunig, Kriegel, Ng, & Sander, 1999). OPTICS is an extension of DBScan (Density-Based spatial clustering of applications with noise) to hierarchical clustering (Ankerst, Breunig, Kriegel, & Sander, 1999).

The advantage of this method is to assigns each data object a degree of being an outlier. However, this method only assigns single objects outliers while ignoring cluster-based outliers

(Duan, Xu, Liu, & Lee, 2009).The formula concept of LOF requires observing some definition related to the symbols described in Table 1:

Table 1
Definition of important Symbols

Symbol	Meaning
x_s, y	Denote objects in a dataset
d(y, z)	Denote the distance between object y and x_s
R	Used for a set of objects
d(y,R)	Denote the minimum distance between y, and object z. since d(y,R) = min{d(y, x_s) | $x_s \in$ R}.
MinPts	A minimum number of objects neighbors
ε	Define a radius around an object, x_s, y ∈D

Definition 1: (local reachability density of an object y): The local reachability density of y is defined as

$$lrd_{minpts}(y) = 1/\left\{\frac{\sum_{x \in N_{minpts}(y)} reach - dist_{minpts}(y, x_s)}{N_{minpts}(y)}\right\} \tag{2}$$

Evidently, the local reachability density of an object y is the average reachability distance based on the MinPts-nearest neighbors of y.

Definition 2: (LOF of an object x): The LOF of y is defined as:

$$LOF_{minpts}(y) = \left(\sum_{N_{minpts}(y)} \frac{lrd_{minpts}(x_s)}{lrd_{minpts}(y)}\right) / \left|N_{minpts}(y)\right| \tag{3}$$

The main goals of using this method are, to determine; (1) how each object in one cluster isolates from their neighbors; (2) how the object in one cluster overlaps based on its neighbors in the other clusters. The result explains that the lower y's local reachability density is, and the higher the local reachability densities of x's MinPts-nearest neighbours are, the higher is the LOF value of y.

Proposed LOF in RKM. The aim of this paper is to provide method to improve the quality cluster partitions by using LOF method in rough clustering. The improvised method applied Local Outlier Factor in Rough K-Means algorithm (LOFRKM) described in algorithm as follows:

Input: K Numbers, ε, MinPts, Threshold.

Output: rough Clusters.

Step 0: Initialization.

- Determine the initial means (e.g., randomly or maximum distance between means).
- Assign each object Xn to the corresponding upper approximation of its nearest mean.

Step 1: Compute the new means as follows:

$$M_k = \left(\sum_{xiIRi)}^{n} \frac{Xi}{\overline{R}i} \right) / \left(\sum_{xiIRi)}^{n} \frac{1}{\overline{R}i} \right)$$

Step 2: Assign the objects to the approximations:

- Assign each Object x to initial upper approximation.
- For each Object in the each Cluster:
    - o Determine if data object detected as outlier or not by LOF threshold as defined below:
        - If the ratio of the LOF value is higher and the MinPts Overlap, then assign data object to an upper approximation of those clusters belong to.
        - Else assign data object to lower approximation of cluster.

Step 3: Check convergence of the algorithm.

- If the algorithm has not converged continue with Step 1.
- Else STOP.

RESULTS

Three experiments were conducted based on the data gathered from synthetic and the UCI Machine Learning Repository datasets.

Data sets

Synthetic data set: Suppose 10 data objects (Figure 5a. Shown the two dimensional objects in graphical space) as follows:

p1= (0.1, 0.0), p2=(0.0, 0.1), p3=(0.1, 0.2), p4=(0.2, 0.0), p5=(0.3, 0.4), p6=(0.6, 0.5), p7=(0.7, 0.8), P8=(0.8, 0.8), p9=(0.8, 0.7), p10=(0.9, 0.8).

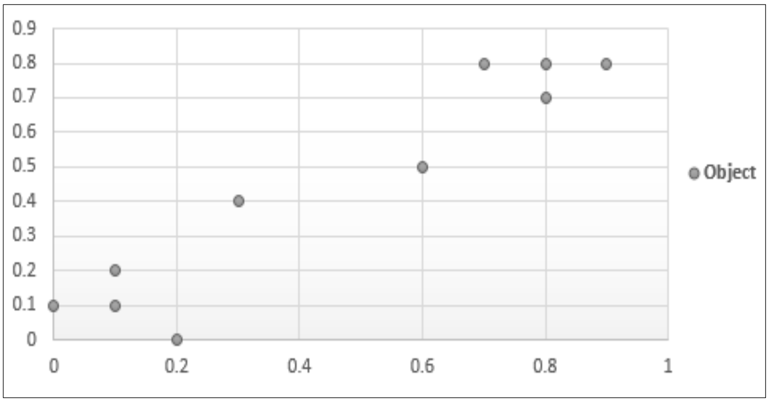

Figure 5a. 10 point two dimensional in graphical space

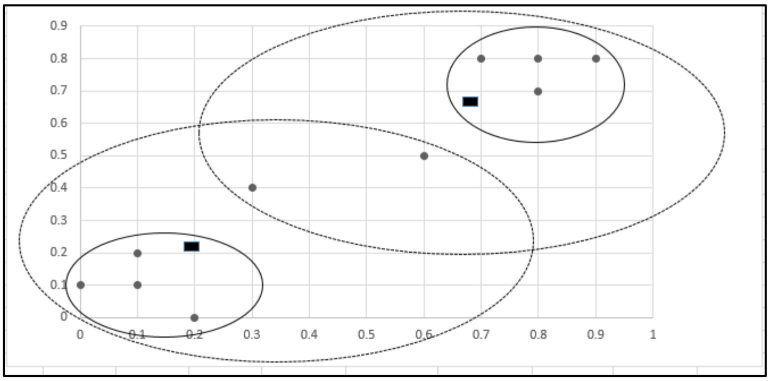

Figure 5b. Two LOFRKM Data Cluster

Synthetic experiment was conducted with two assumptions. First, by using parameters of MinPts=2, k=2 and ε=0.3. The estimated results of Lrds' are given in Table 2a and LOF's are given in Table2b.

Table 2a
Shown the Object Lrd (ε=0.3, MinPts=2)

lrd(p1)	lrd(p2)	lrd(p3)	lrd(p4)	lrd(p5)
7.0721	8.2850	8.2850	5.4794	3.5360
lrd(p6)	lrd(p7)	lrd(p8)	lrd(p9)	lrd(p10)
3.5360	8.2850	7.0721	8.2850	8.2850

Table 2b
Shown the Object LOF (ε=0.3 , MinPts=2)

LOF(p1)	LOF(p2)	LOF(p3)	LOF (p4)	LOF(p5)
4.6859	3.999	3.7073	4.9695	7.0771
LOF(p6)	LOF(p7)	LOF(p8)	LOF(p9)	LOF(p10)
9.9241	3.7072	4.6859	3.7072	3.7072

Second, the number of nearest neighbors was changed from 2 to 3. The results of Lrds' are shown in Table 3a and LOF's in Table 3b. The result indicates that, the LOF value is high when the object Lrd is low. In contrast, Figure 5b shown the Data clusters in paragraph based in first possible results (Table 2a and Table 2b) when the threshold=6. In addition, the means (Centroids) are (0.226, 0.226), (0.665, 0.679).

Table 3a
Shown the Object Lrd(ε=0.3,MinPts=3)

lrd(p1)	lrd(p2)	lrd(p3)	lrd(p4)	lrd(p5)
4.4722	5.0968	5.0968	5.0968	3.5360
lrd(p6)	lrd(p7)	lrd(p8)	lrd(p9)	lrd(p10)
3.5360	6.7965	5.6511	6	6.7965

Table 3b
Shown the Object LOF (ε=0.3, MinPts=3)

LOF(p1)	LOF(p2)	LOF(p3)	LOF(p4)	LOF(p5)
10.2568	8.6322	8.6322	8.6322	12.6463
LOF(p6)	LOF(p7)	LOF(p8)	LOF(p9)	LOF(p10)
17.5949	8.0942	10.6076	9.5670	8.09421

The real dataset from UCI Machine Learning Repository used is the Wisconsin breast cancer data. The data consists 699 instances with 9 attributes and 2 classes named as benign and malignant. The data set was tested with parameters k = 2, and ε= 0.3, MinPts=3. And the second dataset of Iris Plants consists of 150 random samples of flowers and three types of classes which are Setosa, Versicolor, and Virginica. The nature of data set shows that the first class is very easy to separate from the two other classes. With the parameters k = 3 and ε= 0.2, MinPts=3.

Our paper proposed that algorithm is responsible to detect two types of outliers, which are local and global outliers. In the next section, we did a comparative evaluation indicates that LOF improved the quality of cluster portioning.

Clustering Evaluation

In this evaluation, we used two quality measure indexes to produce a quality score for individual clusters. First, the Davies–Bouldin Index (DB Index) is an internal evaluation scheme, where compact the clusters are compared to the distance between the cluster means(Bouldin, 1979). The second Measure is Dunn index (Bezdek & Pal, 1995) defined as the ratio between the minimum distance between point pairs from different clusters and the maximum distance between point pairs from the same cluster. Crucially, large values of the Dunn's index and low values for the DB index are correspond to good data partitions.

The comparison is presented in Table 4, which lists the evaluation methods against Hard K-Means, and existing algorithm πRKM.

Table 4
Comparative performance of clustering algorithms

Dataset	Cluster Algorithm	T	DB Index	Dunn Index
Breast Cancer Wisconsin	K-Means	-	0.7643	6.6905
	π RKM	1	0.2835	7.0422
		0.8	0.2618	7.1696
		0.6	0.1829	7.9299
		0.3	0.1713	10.203
	LOF-RKM	1	0.2653	7.827
		0.8	0.2470	8.341
		0.6	0.1811	10.352
		0.3	0.0083	25.532
Iris Plants	K-Means	-	0.683	1.121
	π RKM	1	0.465	3.314
		0.8	0.428	3.697
		0.6	0.345	4.306
		0.3	0.223	5.174
	LOF-RKM	1	0.362	3.912
		0.8	0.368	3.985
		0.6	0.327	4.474
		0.3	0.208	5.092

The threshold values of all experiments are shown in the above table. The table depicts the values of the DB Index, Dunn Index for different threshold values on two real data sets. The results reported here with the respect to the DB Index, Dunn Index confirm that both the π RKM and LOFRKM achieve the best result for the threshold between 0.3-0.6 (the overall values of threshold have been normalized between 0 and 1). For these particular values of threshold, the performance of LOFRKM is better than the π RKM.

CONCLUSION

Rough clustering is an effective alternative to hard clustering. The quality of clustering partitions has been improved in detecting outliers with the use rough clustering approach. RKM algorithm is directly derived from hard K-Means of proposed properties from the rough set approach. This successful idea has received acceptance and adaption in many application domains. Unfortunately, the need still exists as to find a suitable method to detect outliers in rough clustering partitioning approach. In this paper, we proposed the formulation of LOF and applied it in the rough clustering partitioning algorithm. The results are provided based on synthetic and real datasets. Based on our experiments, we found that, the inclusion of LOF in RKM shows convincing result. And, thus it is evident that RKM can be used to detect outliers in rough clustering partitioning approach. In the future work, we will study to refine the RKM algorithm to be more suitable based on the idea of overlapping clusters.

REFERENCES

Ankerst, M., Breunig, M. M., Kriegel, H., & Sander, J. (1999). OPTICS: Ordering points to ddentify the clustering structure. *ACM SIGMOD International Conference on Management of Data, 2*8(2), 49–60.

Berkhin, P. (2006). A survey of clustering data mining. *Grouping Multidimensional Data*, (c), 25–71.

Bezdek, J. C., & Pal, N. R. (1995). Cluster validation with generalized Dunn's indices. *Proceedings of Second New Zealand International Two-Stream Conference on Artificial Neural Networks and Expert Systems*, 190–193.

Bouldin, D. L. D. A. D. W. (1979). A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence,* (2), 224–227.

Breunig, M., Kriegel, H., Ng, R., & Sander, J. (1999). Optics-of: Identifying local outliers. *Principles of Data Mining and Knowledge Discovery*, 262–270.

Duan, L., Xu, L., Liu, Y., & Lee, J. (2009). Cluster-based outlier detection. *Annals of Operations Research, 168*(1), 151–168.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters, 31*(8), 651–666.

Lingras, P. (2007). Applications of Rough set based K-Means, Kohonen SOM, GA Clustering. *Trans. Rough Sets,* (VII), 120–139.

Lingras, P., & Peters, G. (2012). Rough sets: Selected methods and applications in management and engineering. *Advanced Information and Knowledge Processing, Springer-Verlag London Limited 2012*, 129–138.

Lingras, P., & West, C. (2004). Interval set clustering of web users with rough K-means. *Journal of Intelligent Information Systems, 23*(1), 5–16.

Pawlak, Z. (1982). Rough sets. *International Journal of Computer and Information Sciences, 11*(5), 341–356.

Peters, G. (2006). Some refinements of rough k-means clustering. *Pattern Recognition, 39*(8), 1481–1491.

Peters, G. (2014). Rough clustering utilizing the principle of indifference. *Information Sciences*, *277,* 358–374.

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*(2), 307–322.