International Journal of Engineering & Technology, *7 (2.15) (2018) 41-44*

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research Paper

A hybrid model of ordinal ranking-based clustering using G+Rank K-Means

S. Suhailan [1]*, S. Abdul Samad [2], MA. Burhanuddin [2], M. Makhtar [1]

[1,4] *Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin, 22200, Terengganu, Malaysia*
[2] *Faculty of Information and Communications Technology, Universiti Teknikal Malaysia Melaka, 76100, Melaka, Malaysia*
**Corresponding author E-mail: suhailan@unisza.edu.my*

Abstract

K-Means is a clustering technique that maps object features onto multidimensional coordinates and groups them based on location closeness. However, measuring closest distance can be doubtful when ranking representation of ordinal scale objects are not considered. Thus, distribution of objects in a cluster may violate ranking representation. For example, a same-rank object may be grouped into different clusters. To address this issue, an enhanced of K-Means algorithm is proposed to achieve better and meaningful result of ranking-based clustering. It is based on integration of ranking algorithm that sort objects into ranking list which also representing object closeness based on its nearby location. A new additional step in K-Means is proposed in reassigning unaligned K-Means nearest objects using ranking attribute that eventually accelerates the clustering process. AHP ranking algorithm is integrated into K-Means in achieving a ranking-based cluster. This enhancement was evaluated on three ordinal datasets covering 67 Java programs, 92 students' marks on computer architecture subject and 456 EUFA's football club coefficient ranking list. The results show that by integrating ranking algorithm in K-Means as proposed in G$^+$Rank K-Means, a rank cluster representation has been successfully achieved. The purity value that represents the correctness against certain group classification has also increased.

Keywords: *K-Means, Ranking-Based Clustering*

1. Introduction

Clustering is a technique to automatically group set of objects into unclassified groups based on their features similarity (i.e. closeness). The unclassified groups need to be analyzed by an expert to define and conclude the new findings. Meanwhile, ranking is a technique to order set of objects into certain ranking representation based on ordinal-scaled value. There are many application objects that contain the ordinal features, however can be ranked due to the lack of finite group's definition. For example, lines of codes, speed and intervals, height and weight, are among ordinal features that do not have fixed range that can be used as the definition in recognizing certain groups. This is where clustering and ranking need to be combined so that the result can be represented into a meaningful cluster. These objects can be sorted to represent certain ranks on certain minimum and maximum scale, and then clustered them based on their nearest scale distance.

However, clustering and ranking are often viewed as two independent techniques [1] and are used separately for different purposes. For instance, clustering such as K-Means algorithm is initially performed to minimize the objects (i.e. alternatives) before they can be further processed in the AHP ranking [2-4].

Looking at the clustering process of K-Means, objects are grouped based on their location distance closeness. Such location aspect does not consider the context of the ordinal scale representation. As a result, the cluster may not contain objects with accurate rank representation. This will obstruct the result to be accurately identified as a higher or lower rank cluster. Most of the current researches are focusing on K-Means towards getting a better accurate result by considering initial centroids. However, there is lack of research that investigating on clustering of ordinal features using K-Means algorithm to achieve better and meaningful result for ranking-based clustering. Many previous studies have only focused on integrating the clustering and ranking separately [5–8] which does not imply on ranking-based clustering output effectiveness.

2. Related Works

Most of clustering algorithm such as K-Means, is more concern with measuring closest distance of objects region based on their multiple features space of coordinate location [9-10]. However, considering ranking based application context, ranking level among objects are represented based on their ordinal features. Thus, when clustering is applied, the result is targeted to be aligned with such predefined ranking consideration. Unfortunately, using standard K-Means algorithm, certain objects that represent same rank may not be grouped together because each objects is measured based on closest distance to the centroids rather than among themselves. As clustering result is influenced by initial centroids selection [11], they need also to be targeted towards relevant points of ranking-based cluster representation. Many suggestions have been made on initial centroids enhancement. However, to our extent of knowledge, initial centroids configuration in representing meaningful ranking context is yet to be explored. Different selection of initial centroids may yield to different clustering ranking results due to local minima convergence.

Ranking consideration in clustering is proposed through RankClus [12]. It improves the quality of clustering result by automatically assigning new object feature with calculated rank for each cluster. It starts by generating the clusters and evaluating the rank of each cluster based on individual objects' rank distribution within it. All the objects in the clusters are then assigned with their new rank cluster point. Finally, these objects are re-clustered again by considering the new rank cluster feature for each objects. Thus, the accuracy of clustering is improved by considering the ranking as part of feature. However, this technique may change the objects ranking sensitivity as the rank of cluster are calculated based on initial cluster result which are not yet taking rank representation in the clustering process.

Pei et.al [13] proposes ComClus that is able to calculate centroids based on maximized posterior probability. The posterior probabilities of objects can be used to rank the objects within a cluster. Thus, unimportant objects can be filtered and possibly reassigned to another co-efficient cluster. However, this technique is only limited to networked object that requires relations information on the objects to be available (i.e. dependent features). In certain application, object relation does not exist such as mark of a student does not influence marks for other students.

On the other hand, ComClus has made more sense to develop semi-unsupervised K-Means in which prior knowledge on some labeled data can be used to influence the cluster result towards certain targeted application context [14]. For example, Al-Harbi [15] combines classifier algorithm to create a semi-unsupervised K-Means that can group feature with same label into a same cluster. As main concern in this study is on ranking-based clustering, ranking algorithm is more suitable in providing prior objects information for a semi-unsupervised K-Means clustering result. In this proposed algorithm, AHP are integrated into K-Means algorithm to guide object ranking in generating ranking-based clustering result. However, other ranking algorithm such as weighted-sum model (WSM) or Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) by Hwang may be used based on user specific requirement.

G+Rank K-Means algorithm integrates ranking algorithm to assign objects with ranking attribute that later will be used in guiding K-Means towards achieving better ranking-based clustering representation. It is extended algorithm based on GRank K-Means [16] algorithm that propose new centroid initialization and cluster output reassignment in considering the ranking attribute. The steps are as the following:

2.1. AHP Ranking

This step is needed when the ranking information is not yet established among the objects. Object ranking can be calculated using AHP algorithm that made a relative important comparison between each pair of objects through a comparison matrix (a_{ij}) on each ordinal feature-d. The comparison can be done automatically by calculating the ratio of two different objects on each of the ordinal feature-d. Equation 1 is used to fill up the diagonal and upper triangular of comparison matrix.

$$a_{ijd} = \frac{X'_{id}}{X'_{jd}} \; ; i = 1 \text{ to } N, j = i \text{ to } N \tag{1}$$

Meanwhile, the lower triangular matrix on each object-i is filled up by using equation 2.

$$a = \frac{X'_{jd}}{X'_{id}}; \; i = 2 \text{ to } N, j = 1 \text{ to } (i - 1) \tag{2}$$

This comparison matrix needs to be normalised by dividing each objects-*i* with the summation of their column value as shown in equation 3.

$$A'_{ijd} = \frac{A_{ijd}}{\left(\sum_j^N A_{ijd}\right)} \; ; \; j = 1 \text{ to } N \tag{3}$$

Then, the ranking of object-*i* is aggregated from the normalised comparison matrix using equation 4 by considering the average of C_{ij} in each feature-d where W_d is feature-d weighting and D is the total features.

$$R_i = \sum_d^D \left[W_d * \frac{\sum_j^N A_{ijd}}{N} \right] / D \; ; j = 1 \text{ to } N \tag{4}$$

Using this automated pair-wise ratio calculation, a 100% of consistency is achieved.

2.2. Initial centroids selection

The initial centroid is proposed by considering the object's rank (R_i) so that the clustering process will converge to a cluster with a rank representation. Features of an object (X_i) that has the highest rank will be selected as the initial centroid for the cluster-0, following the second highest rank for the cluster-1, the third highest rank for the cluster-2 and the fourth highest rank for the cluster-3. The formula to select the initial centroids is shown in the equation 5 where k = 0 to K and $\max_k R_i$ is the highest ranking object-i in the descending order. m_k is representing centroid ranking order from the highest rank (m_k) to the lowest rank (m_K).

$$m_k = \left\{ X_i \mid X_i \text{must be in} \max_k R_i \text{ and } R_i > R_{k+1} \right\} \tag{5}$$

2.3. Nearest cluster assignment

Clustering process begins by measuring each object distance on each centroid (m_k) using equation 6.

$$S_{ik} = \min_s \sqrt{\sum_d^D (X_{id} - m_{kd})^2} \tag{6}$$

where S_{ik} is set of the object in cluster-k, k= 0 to K and d is a feature. The objects will be assigned to a cluster where they have the closest distance to the centroid.

2.4. Ranking-based re-clustering

Once the nearest distance to each cluster is completely assigned, this new additional step to guide the result towards object ranking's sensitivity is introduced. This step tries to eliminate inconsistency of ranking-based clustering when using the normal distance measurement that does not assess the object's ranking information. In this step, all objects in a higher cluster but have a lower rank than the highest object rank in a lower cluster will be re-assigned to the lower cluster. This is to ensure the rank cluster to contain only a set of rank objects that does not exceed the rank of the objects in the higher cluster as shown in equation 7.

$$S_{ik} = \left\{ X_i \mid \min_k R_i \leq X_i < \max_k R_i \, , \; R_i \in S_{ik} \right\} \tag{7}$$

where S_{ik} is set of the object in cluster-k, $\min_k R_i$ and $\max_k R_i$ is the lowest-rank and highest-rank of the object in S_{ik}.

2.5. Centroids updates

This is the final step where once the objects have been re-assigned, the centroid for each cluster needs to be re-calculated using equation 8.

$$m_{kd} = \sum_i^M X_{ikd} / M \tag{8}$$

where M is the total of objects in cluster-k, k = 0 to K and d=0 to D. This step is to ensure that all objects that currently assigned to a cluster definitely belong to that cluster (i.e. nearest to its new assigned centroid) and far away from other clusters. If there is an object that turns out to be nearer to another centroid, then this object needs to be reassigned to the nearest cluster. Thus, iteratively, the whole process cycle starting from step (3.3) to (3.4) needs to be repeated until there are no changes to the centroids in all clusters.

3. Experimental Result

Experiments were executed using three datasets; A[17], B[18] and C[19]. Dataset A consists of 67 Java's computer programs that were ranked based on their solution correctness. Dataset B consists of 456 EUFA football's club coefficient ranking result. Dataset C consists of 92 subject's marks of Computer Organization and Architecture. Four clusters (e.g. c0, c1, c2 and c3) were generated for dataset A and B; and three clusters (e.g. c0, c1 and c2) were generated on dataset C using basic K-Means and G$^+$Rank K-Means. Figure 1 shows the clustering result using both algorithms for dataset A, B and C.

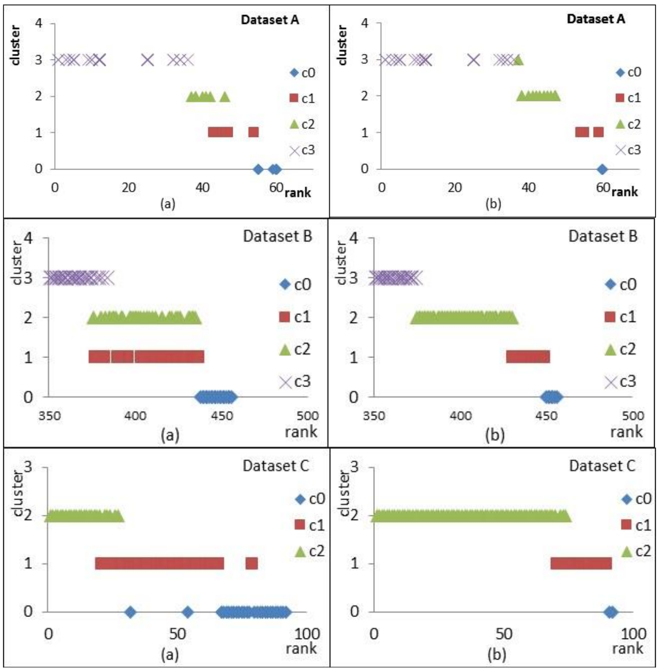

Fig. 1 Ranking-based cluster (a) Basic K-Means and (b) G$^+$Rank K-Means

In Figure 1 (a) where clustering was using the basic K-Means of dataset A and B, some of the objects in a lower rank cluster of c2 have a higher rank than objects in a higher rank cluster of c1. The same result can be observed in dataset C where there are objects of lower ranked that were clustered in a higher cluster, c0. Thus, the clustering result was not clearly separating clusters' member based on their rank attribute. In contrast, when using the G$^+$Rank K-Means, clusters' members were grouped along with their rank's sequence as shown in Figure 1 (b). The result of the experiment shows that the objects were clustered along with the ranking consideration from the highest rank cluster of c0 at the bottom of the graph to the lower consecutive rank cluster (c1, c2 and c3 on dataset A,B; and c1 and c2 on dataset C) from the bottom up.

In order to validate the clustering result in respect to the true label of ranking objects, purity value was used. The cluster results were benchmarked with the targeted ranking group classification as 0-29% (t1), 30-54% (t2), 55-74% (t3) and 75-100% (t4) for dataset A. Table 1 lists the count number of object ranking classification in each clustering result on the dataset A and B.

Table 1: Purity value on Dataset A and B

	Purity	Cluster	Classification (Rubric's percentages)				max
			t1	t2	t3	t4	
Dataset A							
Basic K-Means	0.77612	c0	0	1	2	10	10
		c1	0	1	9	1	9
		c2	0	3	3	1	3
		c3	4	30	2	0	30
G$^+$Rank K-Means	0.79104	c0	0	0	1	7	7
		c1	0	1	1	4	4
		c2	0	4	12	1	12
		c3	4	30	2	0	30
Dataset B							
Basic K-Means	0.9363	c0	0	0	13	5	13
		c1	0	2	0	0	2
		c2	27	24	0	0	27
		c3	384	0	0	0	384
G$^+$Rank K-Means	0.9407	c0	0	0	1	5	5
		c1	0	7	12	0	12
		c2	37	19	0	0	37
		c3	374	0	0	0	374

The result has proven that the proposed methods of GRank K-Means have a better purity value (0.79104) as compared to the random initial centroids of the basic K-Means (0.77612). Thus, the combination of proposed *initial centroid selection* and *ranking-based re-clustering process* enhancement have contributed better accuracy of a ranking-based clustering result. This result is congruent with other researches of the related work which applying semi-supervised clustering to produce better clustering representation.

Meanwhile, for dataset C, the cluster results were benchmarked with the targeted ranking group classification of 0 - 39% (t1), 40 - 74% (t2) and 75 - 100% (t3). Table 2 lists the count number of object ranking classification in each clustering result on the dataset C.

Table 2: Purity value on Dataset C

	Purity	Cluster	Classification (Total mark's percentages)			Max
			t1	t2	t3	
Basic K-Means	0.9565	c0	0	25	2	25
		c1	0	40	0	40
		c2	2	23	0	23
G$^+$Rank K-Means	0.9783	c0	0	0	2	2
		c1	0	24	0	24
		c2	2	64	0	64

4. Experimental Result

This paper has argued that basic K-Means does not put much consideration on ranking-based clustering even for same scale of ordinal features involved. AHP was chosen as a ranking algorithm in guiding K-Means algorithm towards ranking-based clustering consideration. Practical guidance in integrating these ranking algorithm for K-Means clustering process are described in stages. New additional step in K-Means algorithm is proposed to reassign any misaligned object closeness due to centroid constraint by using ranking information consistency guidance. Centroid initialization method is also proposed based on top consecutive ranking objects to minimize bad local minimum convergence of K-Means. The proposed methods are experimented on real data sets that consists of two ordinal features that carry the same scale. The comparison results show significant result that ranking-based clustering can be improved by integrating ranking algorithms to guide K-Means in identifying closest objects.

Acknowledgments

This work has been partially supported by Ministry of Education (MOE) Malaysia under SLAI scholarship.

References

[1] Wang, R., Chen, J., Yu, P. S. & Wu, B. Ranking-based Clustering on General Heterogeneous Information Networks by Network Projection. in *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management* 1, 699–708 (ACM Press, 2014).

[2] Rad, A., Naderi, B. & Soltani, M. Clustering and ranking university majors using data mining and AHP algorithms: A case study in Iran. *Expert Systems with Applications* 38, 755–763 (2011).

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

[4] Singh, A., Yadav, a. & Rana, A. K-means with Three different Distance Metrics. *International Journal of Computer Applications* 67, 13–17 (2013).

[5] Azdnia, A. H., Ghadimi, P. & Aghdam, M. M. A Hybrid Model of Data Mining and MCDM Methods for Estimating Customer Lifetime Value. in *Proceedings of the 41st International Conference on Computers and Industrial Engineering* 80–85 (2011).

[6] Kumar, K. & Kumanan, S. Decision Making in Location Selection: An Integrated Approach with Clustering and TOPSIS. *The IUP Journal of Operations Management* 11, 7–20 (2012).

[7] Dong, B., Gao, P., Wang, H. & Liao, S. Clustering Human Wrist Pulse Signals via Multiple Criteria Decision Making. in *Proceedings of the 26th International Conference on Tools with Artificial Intelligence* (ed. IEEE) 243–250 (2014). doi:10.1109/ICTAI.2014.44

[8] Bai, C., Dhavale, D. & Sarkis, J. Expert Systems with Applications Integrating Fuzzy C-Means and TOPSIS for performance evaluation : An application and comparative analysis. *Expert Systems With Applications* 41, 4186–4196 (2014).

[9] Chormunge, S. & Jena, S. Evaluation of Clustering Algorithms for High Dimensional Data Based on Distance Functions. in *Proceedings of the 2014 International Conference on Information and Communication Technology for Competitive Strategies* Article No. 78 (ACM Press, 2014).

[10] Poomagal, S. & Hamsapriya, T. Optimized k-means clustering with intelligent initial centroid selection for web search using URL and tag contents. in *Proceedings of the International Conference on Web Intelligence, Mining and Semantics - WIMS '11* Article No. 65 (ACM Press, 2011). doi:10.1145/1988688.1988764

[11] Erisoglu, M., Calis, N. & Sakallioglu, S. A new algorithm for initial cluster centers in k-means algorithm. *Pattern Recognition Letters* 32, 1701–1705 (2011).

[12] Sun, Y. *et al*. RankClus : Integrating Clustering with Ranking for Heterogeneous Information Network Analysis. in *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology* 565–576 (ACM Press, 2009).

[13] Pei, J., Tseng, V. S., Cao, L., Motoda, H. & Xu, G. in *Advances in Knowledge Discovery and Data Mining* 583–594 (Springer Berlin Heidelberg, 2013).

[14] Miyamoto, S., Yamazaki, M. & Hashimoto, W. Fuzzy Semi-supervised Clustering with Target Clusters Using Different Additional Terms. in *Proceedings of the International Conference on Granular Computing, 2009, GRC '09* 1, 444–449 (IEEE, 2009).

[15] Al-Harbi, S. H. & Rayward-Smith, V. J. Adapting k-means for supervised clustering. *Applied Intelligence* 24, 219–226 (2006).

[16] Suhailan S. *et al*. Targeted Ranking-Based Clustering Using AHP K-Means. *International Journal of Advance Soft Computing and its Application* 7(3), 100-113 (2015)

[17] Suhailan, S. (2017). Dataset A. [online] figshare. Available at: https://doi.org/10.6084/m9.figshare.5216368 [Accessed 18 Jul. 2017].

[18] Suhailan, S. (2017). Dataset B. [online] figshare. Available at: https://doi.org/10.6084/m9.figshare.5216377 [Accessed 18 Jul. 2017].

[19] Suhailan, S. (2017). Dataset C. [online] figshare. Available at: https://doi.org/10.6084/m9.figshare.5216371 [Accessed 18 Jul. 2017].