Detecting Cluster Numbers based on Density Changes Using Density-index Enhanced Scale-Invariant Density-based Clustering Initialization Algorithm

Onapa Limwattanapibool [1] *and Somjit Arch-int* [2]
[1] Department of Computer Science, Faculty of Science, Khon Kaen, University, Thailand
[2] Department of Computer Science, Faculty of Science, Khon Kaen, University, Thailand
onapa55@hotmail.com, somjit@kku.ac.th

Abstract

Despite high accuracy, K-means relies mainly on the determination of the suitable number of clusters. To cope with, it is hypothesized that in a dataset region with high density tends to be a cluster. The present study is based on Scale-invariant density-based clustering initialization, in which a cluster numbers is derived from density change analysis or density distribution analysis. However, the density calculation under this approach is based on the number and volume of data, which may result in inaccuracy for cluster detection. Thus, the objective of this study was to improve the performance of Scale-invariant density-based clustering initialization to detect the appropriate cluster numbers and initial cluster centers. We proposed a density calculation based on data distance. The density value obtained from the calculation was used as a condition of data division and data merging for cluster detection. According to the experiment, compared to the Scale invariant density-based clustering initialization, the proposed method could detect the cluster numbers and initial cluster centers more equal or closer to the actual number of clusters. In addition, the level of accuracy in clustering was higher than its counterpart.

Keyword: Clustering, K-means, density-based clustering, number of clusters, initial cluster centers.

1. Introduction

K-means is a clustering method by which its high accuracy can be achieved by determination of the appropriate number of clusters; however, it is found difficult to obtain the output[1-4].The hypothesis of this study is that dataset region with high densities has likelihood of be a cluster and that the cluster boundary could be determined by density change and density distribution. A considerable amount of literature has been published on cluster detections by both providing initial cluster center and determining the number of clusters. The following methods lie in the initial cluster center. To illustrate, Cluster Centre Initialization Algorithm (CCIA)[5] makes it possible to avoid local minima trapping, the number of clusters remains necessary to obtain initial cluster center. In addition, the variation coefficient and the correlation coefficient [6] could yield results more accurate than the traditional K-means clustering; nevertheless, the number of clusters remains essential upon cluster determination. Furthermore, Voronoi diagram[7] provides more accurate results compared to those of traditional K-means; however, the obtained initial cluster centers may be an outlier.

Meanwhile, other methods lie in the determination of cluster numbers. To illustrate, disconnectivity[8] is compatible with any clustering algorithm and has low noise sensitivity; however, it causes the insensitive use of computation and fails to provide justification of high

compatibility. Compared to other methods, Artificial Neuron Networks[9] provides relatively accurate results; yet when used with some clusters with uncertain data, this method is reported inconsistency.

Recently, there have been some methods designed to detect both cluster numbers and initial cluster centers, e.g.[10]. His method relies on dynamic determination of cluster numbers and initial cluster centers. The results obtained from this method are comparatively stable; nevertheless, the results depend crucially on the parameter values (Similarity value in Text Similarity Matrix) and compatible with text clustering.

Despite yielding accurate results greater than the traditional K-means clustering, overall the previous methods are unlikely to provide satisfactory results because they may not analyze the data density, the key data for cluster formation, thus resulting in ineffective cluster detection. In addition, Chunsheng Hua, Ryusuke Sagawa and Yasushi Yagi[11] conducted a study on cluster detection by determining the cluster numbers and initial cluster centers. In his method, the cluster boundary is detected through density change or density distribution so called Scale-invariant density-based clustering initialization [11] That is, the change of density is analyzed for cluster boundary as a gap of clusters may form a boundary of clusters. The number and position of clusters could be obtained by merging cluster seeds according to density distribution. The method relies on two parts: division criterion and agglomeration criterion. The first part is intended to divide dataset into regions through density calculation of the regions and its center. The second part is to merge cluster seeds based on density distribution. Scale-invariant density-based clustering initialization[11] could consistently detect the number of clusters and initial cluster centers from the density change by dividing the data and merging process. Such division criterion could yield the initial cluster centers by taking the density change into consideration; meanwhile, the number of clusters could be detected through merging the nearest neighbors cluster seeds, thus increasing the density value of merged clusters. However, this method is based on the density calculation by the number and volume of data, which may provide inaccurate results for decision of dividing and merging regions to detect the cluster numbers. Therefore, the objective of this study was to enhance Scale-invariant density-based clustering initialization method [11] to improve cluster detection performance and eventually increase K-means efficiency. The following sections include the framework of the proposed method, result and discussion.

2. Proposed Method

This section proposes a clustering analysis model to determine the appropriate the number of clusters and initial cluster centers for K-means clustering. This study is based on the scale-invariant density-based clustering initialization[11], This method relies on the calculation of density value, which requires both the number and volume of data; thus, it is likely to suffer from inaccuracy for a decision whether to divide or merge the regions for cluster detection. This study thus addresses those drawbacks by introducing an equation to calculate data density based on data distances. Fig. 1 illustrates the conceptual framework, comprising three parts: (A) Density Calculation, (B) Division Criterion and (C) Agglomeration Criterion. (D) K-means Clustering

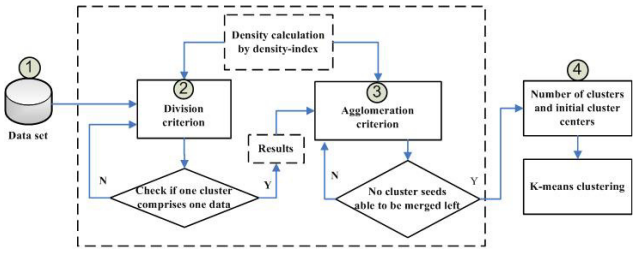


Fig 1 : Framework of density index-enhanced Scale-invariant density-based clustering initialization algorithm

2.1 Density Calculation

In this study, the region density is calculated by density index based on Apinya and Songrit[12], who developed an equation by calculating Eps , β and α for cluster density to evaluate clustering data obtained from DBSCAN. In this study, we improved their equation for region density calculation to form clusters through dividing and merging the region. Thus, the developed equation for region density calculation requires no parameters Eps, β and α . There are density factors taken into consideration: the distance from a considered point to neighboring objects and the number of neighboring objects. It is noted that the density is regarded high only if the distance from the considered point to neighboring objects is short and the number of neighboring objects is high. See the following correlation.

$$\text{Density Index} \propto \frac{N_{neighboring\_objects}}{\sum d_{to\_neighboring\_objects}}$$

Where $N_{neighboring\_objects}$ represent the number of neighbouring objects

$d_{to\_neighboring\_objects}$ represent the distances to neighbouring objects

$\sum d_{to\_neighboring\_objects}$ represent sum of distances to neighbouring objects

According to the above density index correlation, each of the considered points is determined as o_i. If the distance from o_i to neighbouring objects equals zero, the data thereby suffer from division by zero. To solve this problem we propose function f_z, which is obtained from a 2-degree polynomial function related to a 2-dimensional space dataset see (1).

$$f_z(d(o_i,o_j)) = \begin{cases} 1 & ;d(o_i,o_j)=0 \\ d(o_i,o_j) & ;d(o_i,o_j)>0 \end{cases} \quad (1)$$

where $d(o_i,o_j)$ represents the Euclidian distance between objects o_i and o_j. Hence, the density index function of object o_i can be calculated as follows:

$$f_{den_r}(o_i) = \frac{|N_r(o_i)|}{\sum f_z(d(o_i,o_j))}, \quad o_j \in N_r(o_i) \quad (2)$$

Where

$N_r(o_i)$ represents the neighboring object of object o_i in the radius of Eps in the region.

$|N_r(o_i)|$ represents the number of objects in the neighboring object of object o_i in the radius of Eps in the region.

$f_{den}(o_i)$ represents the density index function of object o_i.

The key characteristics $f_{den}(o_i)$ is :

$$\lim_{x->0^+} x < f_{den}(o_i) \le 1$$

The density value of the cluster is derived from the density index function as follows:

$$D_r = \frac{\sum_{i=1}^{n} f_{den_r}(o_i)}{|N_r(o_i)|}, o_j \in N_r(o_i), i \neq j \quad (3)$$

2.2 Division criterion

Division criterion is a method of cluster detection by dividing data. Prior to data division, to avoid time-wasting computation, the dead cluster seeds were eliminated, while the remained seeds were used for data division. The remained seeds were divided into four regions, each of which was similar in size. Individual region was represented by $R_j^i (j=1-4)$ and computed for region density by the density index (see 3.1).

Subsequently, define R_i^{center} of region R_j^i by determination of a center region of region R_j^i, and calculate the center density value of region (d_{center}^i) where the regions (R_j^i) were divided by $d_{center}^i \le d_j^i$ (see figure 1).

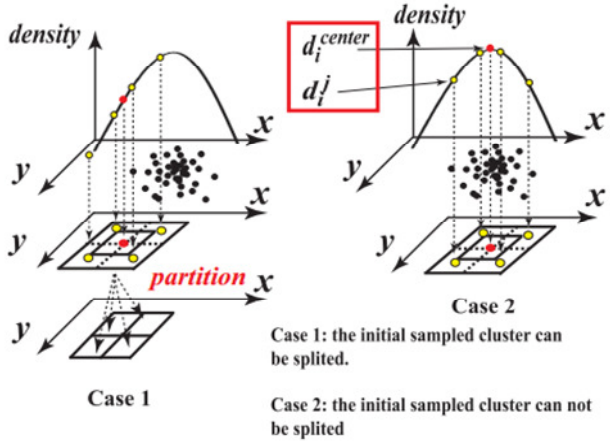


Fig 2 : shows data division in Division Criterion [11]

According to the Fig. 2, there are two cases. The first one shows a region division to develop clusters under $d_{center}^i \le d_j^i$; in the second one, circumstances not applicable under $d_{center}^i \le d_j^i$ is unlikely to be divided. It is noted that the division process will be repeated over and over to obtain one data per cluster.

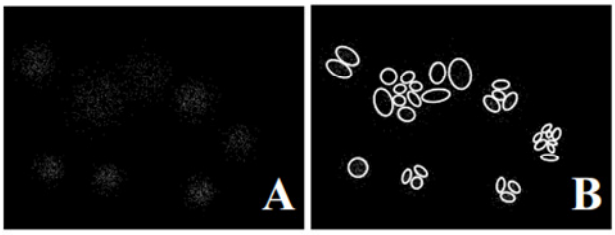
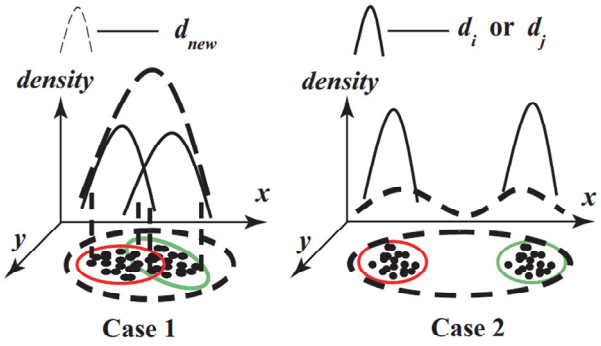


Fig 3 : shows data set and outputs from division criteria[11]

The Fig. 3(figure A) illustrates data set divided by the division criterion (see figure B). According to the output, the gathered data sets are divided into a variety of regions.

2.3 Agglomeration criterion

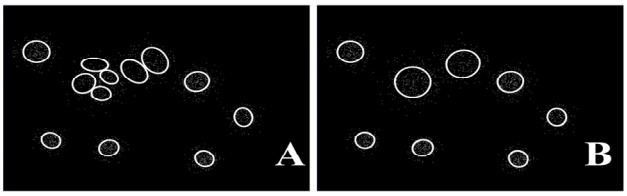
After division criterion, the agglomeration criterion is used to merge the regions of results obtained from division criterion for cluster detection. To start with, the obtained data are plotted into a density histogram. Consider valleys lying between two peaks in the density histogram. We consider the gap(whose density is about zero) between clusters as the cluster boundary. Calculate for density value in each cluster. Eventually, consider the merging of two clusters under $d_{new} \ge \arg\min(d_i, d_j)$. See Fig. 4.



Case 1: two clusters should be merged
Case 2: two clusters should not be merged

Fig 4 : shows data merging in Agglomeration Criterion[11]

According to the Fig. 4, the case one shows two clusters that party overlap each other. To determine d_{new}, obtained from $d_i \cup d_j$, the density of d_{new} is higher that that of either d_i or d_j, which meets the circumstances allowing to merge the clusters. The case 2 shows two clusters with huge gap. The determination of d_{new} shows that the density obtained from $d_i \cup d_j$ causes the value of d_{new} to less than lower than either d_i or d_j where not applicable. The process is repeated until no more cluster seeds are left to merge.



Merge Once **Merge Twice**

Fig 5 : shows the results obtained from cluster merging by division criterion Criterion [11]

According to Fig. 5, A refers to the result derived from Agglomeration Criterion of the clusters divided by Division Criterion in the first time and B referring to the result in the second time, respectively.

2.4 K-means Clustering

The final results by Agglomeration criterion yield the appropriate number of clusters and the initial cluster centers applicable to K-means Clustering for improved performance.

3. Experiment and Evaluation

In this study, six dataset derived from UCI database were used for the experiment to determine the appropriate number of clusters and initial cluster centers for an application with K-means clustering. The results by the proposed method are then compared with those by Scale-invariant density-based clustering [11] (see Table 1).

Table 1: Shows a comparison of results obtained by the proposed method and by Scale-invariant density based clustering

Dataset	Actual #Cluster[a]	Min pts[b]	Accuracy			
			# Clusters[c]	Scale-invariant density-based clustering	# Clusters[d]	Proposed model
Ecoli	8	6	6	60.12	8	65.18
Liver-disorders	2	7	3	49.28	2	55.36
New-Thyroid	3	5	5	51.16	4	77.67
Balance-scale	3	9	3	54.88	3	63.20
Seed	3	8	1	33.33	5	71.43
Iris	3	6	2	66.67	3	89.33

[a] *Actual number of clusters*

[b] Minimum point of clusters

[c] Number of clusters of Scale-invariant density-based clustering

[d] Number of clusters of proposed method

According to Table 1, the proposed method could yield the results more accurate than the control method in all of datasets. The number of clusters detected by the proposed method was equal to the actual number of clusters in four dataset: Ecoli, Liver-disorders, Balance-scale and Iris and close to the actual number in two datasets: New-Thyroid and Seed. Fig. 5 compares the accuracy value of both methods. It is noted that the accuracy shown on the table means an accuracy value of data clustering by considering whether the data are clustered in the actual dataset region; thus, the accuracy value is unnecessarily relevant to the number of clusters.

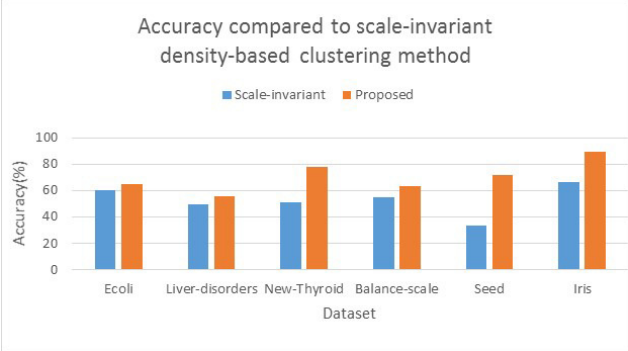


Fig 6 : Comparison of accuracy of K-means obtained from the number of clusters yielded by the proposed method and Scale-invariant density-based clustering

4. Conclusion

In this study, it is hypothesized that dataset region with high density is likely to be a cluster. The cluster could be determined through changes and distribution of data. The findings suggest that when used with K-means, the number of clusters and initial cluster centers obtained from the proposed method has a higher accuracy than that of the scale-invariant density-based clustering. The proposed method could locate a high density region likely to be its cluster for cluster detection through data

distribution. In addition, the proposed calculation using the distance could detect the number of clusters more accurately and closely and yield accuracy value higher than the use of Scale-invariant density-based clustering.

References

[1] M. Xu, *et al.*, "A medical procedure-based patient grouping method for an emergency department," *Applied Soft Computing,* vol. 14, Part A, pp. 31-37, 2014.

[2] K. D. Kaushik H. Raviya, "An empirical comparison of K‐means and DBSCAN clustering algorithm," *Paripex Indian Journal Of Research,* vol. 2, April 2013.

[3] J. Liang, *et al.*, "Determining the number of clusters using information entropy for mixed data," *Pattern Recognition,* vol. 45, pp. 2251-2265, 2012.

[4] K. Liao, *et al.*, "A sample-based hierarchical adaptive K-means clustering method for large-scale video retrieval," *Knowledge-Based Systems,* vol. 49, pp. 123-133, 2013.

[5] S. S. Khan and A. Ahmad, "Cluster center initialization algorithm for K-means clustering," *Pattern Recognition Letters,* vol. 25, pp. 1293-1302, 2004.

[6] M. Erisoglu, *et al.*, "A new algorithm for initial cluster centers in k-means algorithm," *Pattern Recognition Letters,* vol. 32, pp. 1701-1705, 2011.

[7] D. Reddy and P. K. Jana, "Initialization for K-means Clustering using Voronoi Diagram," *Procedia Technology,* vol. 4, pp. 395-400, 2012/01/01 2012.

[8] J.-S. Lee and S. Olafsson, "A meta-learning approach for determining the number of clusters with consideration of nearest neighbors," *Information Sciences,* vol. 232, pp. 208-224, 2013.

[9] N. Alp Erilli, *et al.*, "Determining the most proper number of cluster in fuzzy clustering by using artificial neural networks," *Expert Systems with Applications,* vol. 38, pp. 2248-2252, 2011.

[10] P. J. a. L. J. Shao Xiongkai, "A method of dynamically determining the number of clusters and cluster centers," presented at the 2013 8th International Conference on Computer Science & Education (ICCSE), Colombo, Sri Lanka, 2013.

[11] R. S. a. Y. Y. Chunsheng Hua, "Scale-invariant density-based clustering initialization algorithm and its application," presented at the 2008 19th International Conference on Pattern Recognition, Tampa, Florida, 2008.

[12] A. T. a. S. Maneewongwattana, "U-DBSCAN : A density-based clustering algorithm for uncertain objects," presented at the 2010 IEEE 26th International Conference on Data Engineering Workshops (ICDEW 2010), Long Beach, California, 2010.