Journal of Software Engineering

ISSN 1819-4311

Academic Journals Inc.

www.academicjournals.com

Journal of Software Engineering 9 (1): 179-187, 2015
ISSN 1819-4311 / DOI: 10.3923/jse.2015.179.187

Measuring Overlap-Rate in a Hierarchical Approach for Determining the Number of Clusters

[1]Zhiling Hong and [2]Meihong Wu
[1]School of Software, Xiamen University, 361005, China
[2]School of Information Science and Technology, Xiamen University, 361005, China

Corresponding Author: Meihong Wu, School of Information Science and Technology, Xiamen University, 361005, China

ABSTRACT

Determining the number of clusters is one of the most important topic in cluster analysis. The ability of clustering algorithm to distinguish between overlapping clusters is one of the major criteria for evaluating its efficiency. However, cluster overlapping phenomenon is not yet well understood by researchers. This study, a definition was presented on the degree of overlap between two clusters firstly. Based on this definition, an algorithm was developed for calculating the overlap rate. Then it was show that how the theory can be used to deal with the problem of cluster merging in a hierarchical approach to clustering and gave an optimal number of clusters automatically. Furthermore, comparison was made between the proposed new method and the previous related works. Finally, the experimental results demonstrate the effectiveness of the overlap rate measuring method and the new hierarchical clustering algorithm.

Key words: Cluster validity, overlap rate, cluster merging, hierarchical

INTRODUCTION

Data clustering is the process of grouping data into clusters so that the objects within a cluster are highly similar and the objects in different clusters are highly dissimilar. Automatically determining the number of clusters is one of the most important topics in cluster analysis and still is an open problem.

A common approach for determining the number of clusters is an iterative trial and error process based on a cluster validity index (Sun *et al*., 2004; Zhou *et al*., 2014). Based on cluster validation, this approach consists of running a partition-based clustering algorithm such as the K-Means (Krishna and Murty, 1999; Krishnasamy *et al*., 2014) or FCM (Bezdek, 1981) for a range of numbers of clusters, testing the results according to a validity index defined as a function of the trade-off between within-cluster compactness and between cluster separation and choosing the number that optimizes the validity index. Their performance depends on whether the data set contains well separated clusters, or in other words, whether and how the clusters overlap each other. In addition, it can be very time-inefficient and often fails when some clusters deviate too much from a spherical.

In a given a data set, almost all clustering algorithms are able to distinguish well separated clusters. However, for data sets with overlapping clusters, results often unpredictable. One of the main reasons for this problem is that many algorithms fail to distinguish partially overlapping clusters. It is thus necessary to be able to precisely measure the degree of overlap for deciding

whether overlapping clusters are distinguishable from each other. For example, in image processing, the degree of overlap between two clusters in a color image measures the similarity of the objects.

In this study, a definition on the degree of overlap between two clusters firstly is presented. Based on this definition, an algorithm for calculating the overlap rate was developed. Then, it was shown how the theory can be used to deal with the problem of cluster merging in a hierarchical approach to clustering and provide a natural way of determining the number of clusters automatically. Finally, experimental results demonstrated the effectiveness of overlap rate measuring method and the new hierarchical clustering algorithm.

MATERIALS AND METHODS

Degree of overlap in clusters: A set of n entities forming a k-mode array can be presented as:

$$X = \{X_1, X_2, \ldots, X_n\}$$

where X_i is a vector of dimension d. Each X_i has been partitioned into one of k clusters. For the sake of simplicity, the theory will be introduced in the 2-D case, d = 2 and all the results hold in the multidimensional case.

The methods used to measure the similarity (degree of overlap) between two clusters can be grouped into three main types: Fuzzy set, geometric and probabilistic methods.

Fuzzy set methods are based on the result of fuzzy c-means algorithm and the measure of overlap indicates the degree of overlap between fuzzy clusters (Kim *et al*., 2004). For the fuzzy cluster, the overlap function f(μ) at a given membership degree μ between two fuzzy clusters $\tilde{F}_p$ and $\tilde{F}_q$ is defined as:

$$f(\mu : \tilde{F}_p, \tilde{F}_q) = \sum_{j=1}^{n} \delta(x_j, \mu : \tilde{F}_p, \tilde{F}_q) \quad (1)$$

$$\delta(x_j, \mu : \tilde{F}_p, \tilde{F}_q) = \begin{cases} \omega(x_j) & \text{if } \mu_{\tilde{F}_p}(x_j) \geq \mu \text{ and } \mu_{\tilde{F}_q}(x_j) \geq \mu \\ 0.0 & \text{otherwise} \end{cases} \quad (2)$$

where, $\delta(x_j, \mu: \tilde{F}_p, \tilde{F}_q)$ determines, if two clusters are overlapped at the membership degree μ for the data point X_j. It returns an overlap value of $\omega(x_j)$ when the membership degrees of the two clusters are both greater than μ.

Geometric methods focus on the geometrical properties of the clusters. Typical similarity measures include minimum distance (single-link) maximum distance (complete-link) average distance (average-link) centroid distance (centroid-link) inner squared distance (ward-link) and so on (Kim *et al*., 2004). All of these distances are based on direct comparison between points in two clusters and do not take into account the distribution properties of each cluster. Some of these methods are designed for edge detection in image processing (Tabbone, 1994) or for generating trusted component data sets (Aitnouri *et al*., 2002).

Probabilistic methods are based on a hypothesis that data has drawn from one of several probabilistic distributions. The most commonly used model is the Gaussian mixture and various distances such as the Mahalanobis distance:

$$D_{mah} = ((\mu_1 - \mu_2)^T \Sigma^{-1}(\mu_1 - \mu_2))^{1/2} \quad (3)$$

which is assuming that they have the same covariance matrix and an extension of it.

In what follows, it was tried, to characterize the overlap phenomenon. In particular, an efficient procedure for verifying whether the two clusters strongly overlap and to compute an overlap rate was derived when they partially overlap.

Definition of the overlap rate: For a given two clusters (C_i, C_j), the numbers of them are (N_i, N_j) separately and the overlap rate (OLR) between them is determined by the ratio of the number of overlap points to that the number of small cluster's points:

$$OLR(C_i,C_j)=\begin{cases} 1 & \text{if } N_{Over\_Region} \geq N_{min} \\ \dfrac{N_{Over\_Region}}{N_{min}} & \text{others} \end{cases} \quad (4)$$

where, $N_{Over\text{-}Region}$ is the number of the overlap points, N_{min} = min (N_i, N_j) is the minimum value of N_i and N_j. OLR falls in [0, 1].

Algorithm for calculating overlap rate based on "Region label": In classical physics, the law of gravity between mass is defined by:

$$F(m_1,m_2) = G\frac{m_1 m_2}{r^2}$$

Inspired by the idea of "Universal gravitation", so every sample was treated as a mass and a set of them form a larger mass which corresponds to a cluster. Because gravitation is inversed ratio to r^2, it is descending when the sample is away from the cluster. Gravitation was simplified into four levels, corresponding to four "Distance regions" R_1, R_2, R_3, R_4, (Fig. 1). The formal definition of determining a sample in what "distance regions" is as following.

So, in given cluster C_i and a sample y_k, the y_k can be determined as it belongs to one of the four "Distance regions" by:

$$RegNum=\begin{cases} R_1 & \text{if } 0 \leq R_{ki} \leq R_{Ci} - S_{Ci} \\ R_2 & \text{if } R_{Ci} - S_{Ci} \prec R_{ki} \leq R_{Ci} \\ R_3 & \text{if } R_{Ci} \prec R_{ki} \leq R_{Ci} + \alpha S_{Ci} \\ R_4 & \text{if } R_{Ci} + \alpha S_{Ci} \prec R_{ki} \end{cases} \quad (5)$$

where, R_{ki} = d (y_k, $\bar{x}$ c_i), $\bar{x}_{ci}$ is the center of cluster C_i and d(y_k, $\bar{x}_{ci}$) is Euclidean distance between y_k and C_i, R_{Ci} is radius of cluster C_i:

$$R_{Ci} = \frac{1}{N_{c_i}}\sum_{i=1}^{N_{ci}} d(x_i, \bar{x}_{c_i})$$

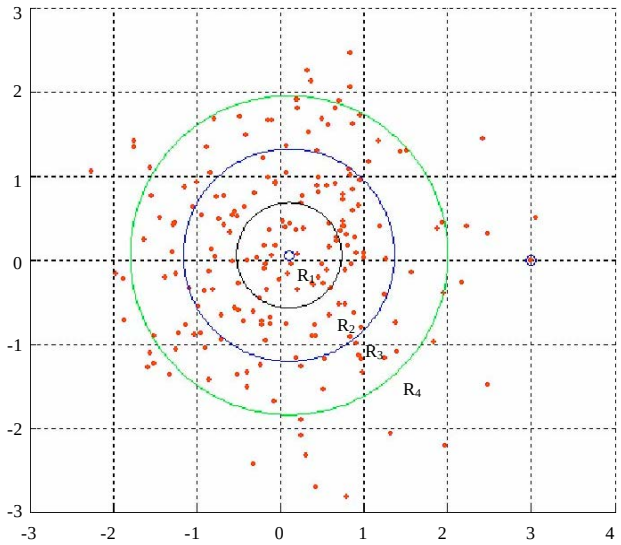

Fig. 1: Four "distance regions"

Table 1: Possible combinations of region numbers for clusters C_i, C_j

C_i	C_j	Overlap point	C_i	C_j	Overlap point
R_1	R_2	1	R_1	R_1	-
R_1	R_3	1	R_1	R_4	0
R_2	R_1	1	R_2	R_4	0
R_2	R_2	1	R_3	R_4	0
R_2	R_3	1	R_4	R_1	0
R_3	R_1	1	R_4	R_2	0
R_3	R_2	1	R_4	R_3	0
R_3	R_3	1	R_4	R_4	0

which is the mean of all the distances. S_{Ci} is the standard deviation of all the distances, α is a parameter to control the probability that a sample falling in the interval $[0, R_{Ci}+\alpha S_{Ci}]$. According to definition 2, the samples was labeled with a region number in a cluster. Given two clusters C_i, C_j, for the entire sample in D, $D = C_i \cup C_j$. Every samples can be labeled with two region numbers, one is for cluster C_i and the other one for C_j. All the possible combinations of region numbers are illustrated in Table 1. "1" denotes that the sample with the combination of R_i and R_j is an overlap point and the combination corres ds to a kind of overlapping region. "0" denotes not. The collection of all the overlapping regions can be visually shown in Fig. 2. In two given clusters C_i, C_j, the number of overlap points which are in the intersection regions could be easily found.

Algorithm 1 (RL_OLR) computes the overlap rate of any two clusters (represented by C_i and C_j), where, α, is a parameter that control the probability a sample that falling in the largest circle. In general, $\alpha = 3$ to ensure that the interval $[0, R_{Ci}+3S_{Ci}]$ will contain about 99% of the samples of this cluster.

Method of the hierarchical clustering based on overlap-rate: Hierarchical clustering approaches generate a nested series of partitions by merging clusters (agglomerative approach) or

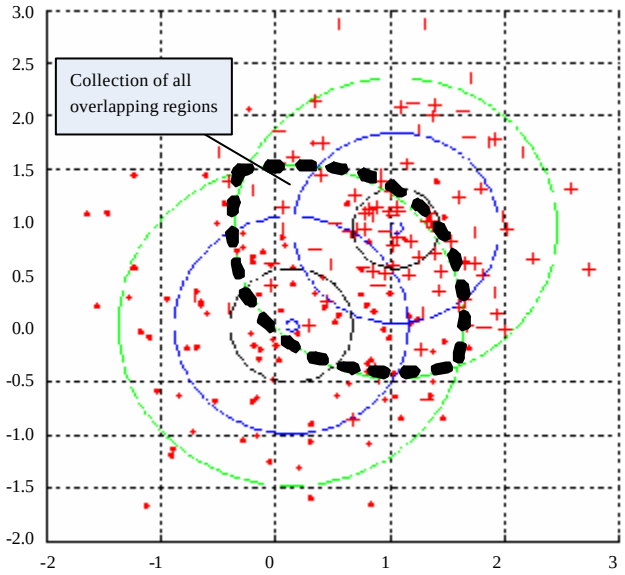

Fig. 2: Collection of all overlapping regions

Algorithm 1: Measuring overlap-rate base on region label (RL_OLR)

Begin

1. Input the parameter α, if α = Null, then α = 3
2. For each sample y_k in D, D = $C_i \cup C_j$
   2.1 Label a region number relative to C_i by Eq. 5
   2.2 Label another region number relative to C_j by Eq. 5
   2.3 Determine whether y_k is an overlap point with the two region numbers according to Table1
3. Statistic all of the overlap points
4. Compute OLR of the two clusters by Eq. 4

End

splitting them (divisive approach), based on a measure of similarity. Measuring the similarity between clusters is the key to hierarchical algorithms. Typical similarity measures include minimum distance (single-link), maximum distance (complete-link) and average distance (average-link) (King, 1967). The OLR was proposed as a measure of similarity. It provides a natural way to merge two clusters.

OLR-based merging strategy in a hierarchical clustering algorithm: Before applying OLR as a measure of similarity for hierarchical clustering, each data X_i should have been partitioned into one of k clusters. The improved Fuzzy C-Means (KFCM) algorithm with Gaussian-Kernel (Zhang and Chen, 2004) to yield a set of clusters which provides good coverage of the data set was applied. The hierarchical clustering techniques can then be applied to the results of the KFCM algorithm.

The algorithm contains two input parameters that can be easily set by the user.

α is the parameter to control the probability that a sample falling in the largest circle. According to our experience, any value between 1 and 3 is a good choice. It can be null to use the default value 3 which ensure that the interval [0, R_{Ci}+3S_{Ci}] contain about 99% of the samples in cluster.

Algorithm 2: Cluster merging using overlap-rate measuring (CM_OLR)

Begin

Step 1: Gaussian-Kernel based KFCM algorithm to yield a set of clusters

Step 2: Cluster Hierarchical Merging using Overlap-Rate Measuring

1. Enter values for the parameters α and η, set C = K
2. Calculate the overlap rate, OLR_{ij}, between each pair of clusters ($1 \le i < j \le C$) using Algorithm1 RL_OLR and calculate the maximum overlap rate:

$$\text{Max OLR} = \max_{1 \le i < j \le C} \{OLR_{ij}\}$$

3. Select and merge all possible candidate cluster pairs

3.1 Select candidate cluster pairs

A pair of clusters (i, j) is a candidate if its OLR_{ij} satisfies:

$$\text{MaxOLR} - \eta < OLR_{ij} \le \text{MaxOLR}$$

3.2 Merge the selected candidate cluster pairs

A candidate is selected for merging if its OLR_{ij} satisfies:

$$OLR_{ij} = \max_{k \ge j, l \le i} \{OLR_{i,k}, OLR_{l,j}\}$$

4. Update the current number of clusters $C_{current}$
5. If C > 2 and $C_{current} < C$, then $C = C_{current}$ and go to Step 3, otherwise, output the clusters and stop (C is the number of clusters)

End

η is a constant for controlling the merging process. The merging procedure selects all pairs whose OLR value falls within the interval (MaxORL-η, MaxOLR) as candidates for merging. η has some impact on the time efficiency and on the number of final clusters obtained. According to our experience, any value between 0.05 and 0.15 is a good choice. In this study, $\eta = 0.1$ in all tests of the proposed algorithm was used.

RESULTS AND DISCUSSION

Measuring the overlap of clusters in generated data sets: The following examples demonstrated that the overlap-rate measuring method proposed in this study is the better similarity measure than the average distance.

Figure 3a and b both contain artificially generated clusters. The cluster centered at the origin is same in both figures. The other cluster in each of the two figures has the same center, orientation and length on the principal axis. They differ only in the width on the secondary axis. Table 2 illustrates the similarity measuring results between average-distance-based method and OLR-based method proposed in this study.

It was seen that the overlap rates between two clusters in each figure are very different (OLR = 0.763 in Fig. 3a vs. OLR = 0. 432 in Fig. 3b) while the average distance is very similar in both figures (3.578 vs. 3.541). Thus, according to the average distance, the strange conclusion was obtained that two clusters in Fig. 3b are more similar than the clusters in Fig. 3a.

Another advantage of using the overlap rate as a similarity measure, compared to distance measures, is that the value of OLR is normalized between 0 and 1. This makes the values of similarity between clusters more comparable.

Color image segmentation using proposed hierarchical clustering algorithm: Several results of the proposed algorithm in application to color image segmentation were

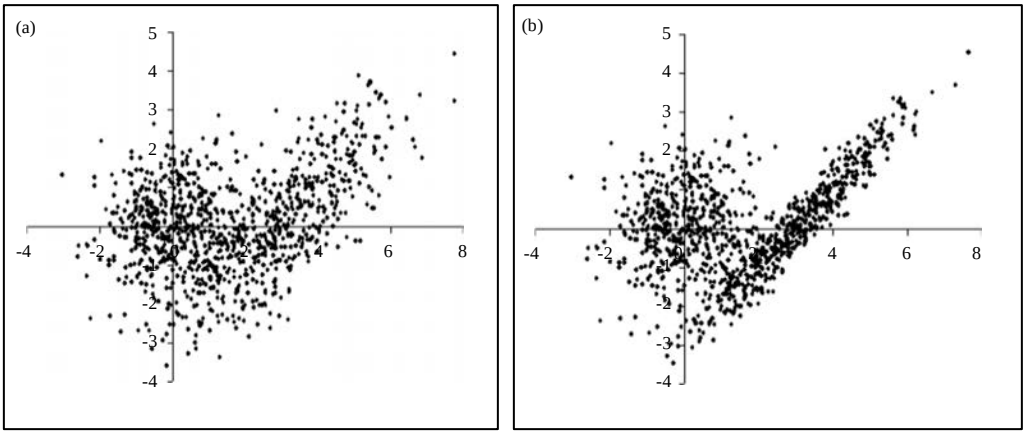

Fig. 3(a-b): Generated clusters for comparing (a) Average-distance-based and (b) OLR-based similarity measures

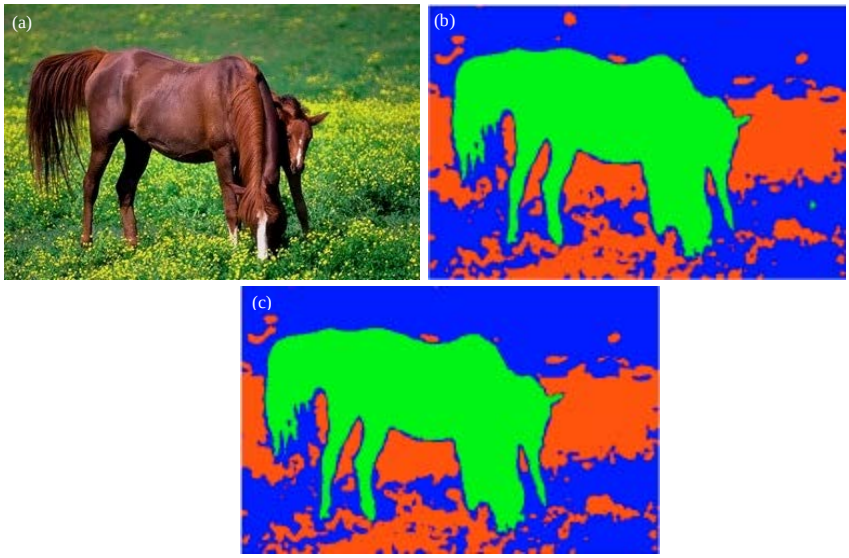

Fig. 4(a-c): Comparison between k-means and CM_OLR algorithm, (a) Original image, (b) Image obtained by applying the k-means algorithm with k = 3 and (c) Image obtained by the new algorithm (the number of clusters which is 3, is obtained by the new algorithm)

reported. In order to show the effectiveness of the algorithm, the pre-processing of each image was limited to minimum. In fact, the transformation to the Lab color coordinate system is applied to the images and only take (a, b) from image as input to the new Algorithm CM_OLR. Throughout the experiments the two input parameters were set as follows: $\eta = 0.1$ and α leave it as default $\alpha = 3$.

For the Horse image, the k-means algorithm (Krishna and Murty, 1999; Krishnasamy *et al.*, 2014) is initialized using the number of clusters obtained by the new algorithm. The results are shown in Fig. 4. By examining the results in Fig. 4b and c, it can be clearly seen that the new

Table 2: Similarity measuring comparison between average-distance-based and OLR-based methods

	Similarity measuring methods	
Comparison of Similarity measuring methods	Average distance	RL_OLR
Figure 3a	3.578	0.763
Figure 3b	3.541	0.432

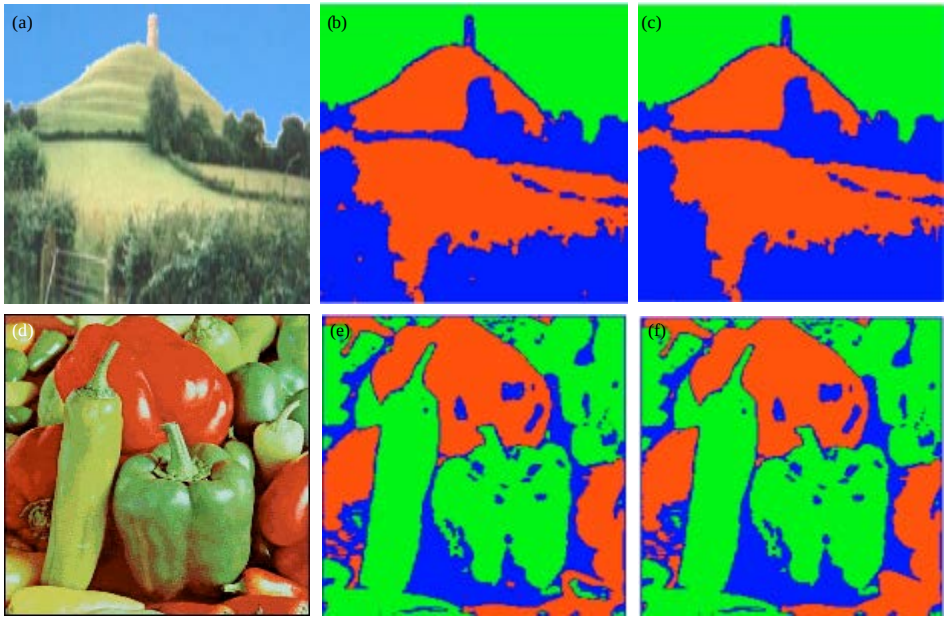

Fig. 5(a-f): Performance of the proposed algorithm, (a, d) Original images (b, e) Image obtained by applying the k-means algorithm with k = 3 separately and (c, f) Corresponding segmented images yielded by the proposed algorithm

algorithm performs well in merging gradually changing pixels within each region and in preserving the natural boundaries between regions. These properties are also reflected in the results with the other two images (Fig. 5).

CONCLUSION

The main contribution of this study is establishment of a theory to explain the phenomenon of overlap and provide a computationally feasible way to calculate the degrees of overlap between two clusters. Specifically, the algorithm RL_OLR has been proposed to considers the effect of cluster, dimension, cluster size, mean, standard deviation, cluster orientation and shape which is much superior to other similarity measuring methods.

Furthermore, in order to determine the number of clusters, OLR was used as a measure of similarity and design a novel hierarchical algorithm CM_OLR that is capable of automatically determining the number of clusters. Comparison between this new method and pervious related works was made and the experimental results demonstrate the effectiveness of the overlap rate measuring method and the new hierarchical clustering algorithm.

ACKNOWLEDGMENT

This study is supported by National Natural Science Foundation of China (31200769), the 2014 Program for New Century Excellent Talents in Fujian Province University and the Open Funding Project of Zhejiang Key Laboratory for Research in Assessment of Cognitive Impairments.

REFERENCES

Aitnouri, E., F. Dubeau, S. Wang and D. Ziou, 2002. Controlling mixture component overlap for clustering algorithms evaluation. Pattern Recognit. Image Anal., 12: 331-346.

Bezdek, J.C., 1981. Pattern Recognition with Fuzzy Objective Function Algoritms. 1st Edn., Plenum Press, New York, USA.

Kim, D.W., K.H. Lee and D. Lee, 2004. On cluster validity index for estimation of the optimal number of fuzzy clusters. Pattern Recognit., 37: 2009-2025.

King, B., 1967. Step-wise clustering procedures. J. Am. Stat. Assoc., 69: 86-101.

Krishna, K. and M.N. Murty, 1999. Genetic K-means algorithm. IEEE Trans. Syst. Man Cybern. Part B. Cyber., 29: 433-439.

Krishnasamy, G., A.J. Kulkarni and R. Paramesran, 2014. A hybrid approach for data clustering based on modified cohort intelligence and K-means. Expert Syst. Applic., 41: 6009-6016.

Sun, H., S. Wang and Q. Jiang, 2004. FCM-Based model selection algorithms for determining the number of clusters. Pattern Recognit., 37: 2027-2037.

Tabbone, S., 1994. Edge detection, subpixel and junctions using multiple scales. Ph.D. Thesis, Institute National Polytechnique de Lorraine France.

Zhang, D.Q. and S.C. Chen, 2004. A novel kernelized fuzzy C-means algorithm with application in medical image segmentation. Artif. Intell. Med., 32: 37-50.

Zhou, K., S. Ding, C. Fu and S. Yang, 2014. Comparison and weighted summation type of fuzzy cluster validity indices. Int. J. Comput. Commun. Control, 9: 370-378.