ON CLUSTER VALIDITY INDEXES IN FUZZY AND HARD CLUSTERING ALGORITHMS FOR IMAGE SEGMENTATION

Moumen El-Melegy, E.A.Zanaty***, Walaa M.Abd-Elhafiez**, and Aly Farag§*
*Electrical Department, Faculty of Engineering, Assiut University, Assiut, Egypt.
**Math. Department, Faculty of Science, Sohag University, Sohag, Egypt.
***Computer Science Dept., Faculty of Comp. and Information System, Taif University, Saudi Arabia.
§Computer Vision and Image Processing Lab., University of Louisville, KY 40292, USA

ABSTRACT
This paper addresses the issue of assessing the quality of the clusters found by fuzzy and hard clustering algorithms. In particular, it seeks an answer to the question on how well cluster validity indexes can automatically determine the appropriate number of clusters that represent the data. The paper surveys several key existing solutions for cluster validity in the domain of image segmentation. In addition, it suggests two new indexes. The first one is based on Akaike's information criterion (AIC). While AIC was devoted to other domains such as statistical estimation of model fitting, it is implemented here for the first time as a validation index. The second index is developed from the well-established idea of cross-validation. The existing and new indexes are evaluated and compared on several synthetic images corrupted with noise of varying levels and volumetric MR data.

***Index Terms*—**clustering, cluster validity, fuzzy clustering, image segmentation.

1. INTRODUCTION

Cluster analysis is based on partitioning a collection of data points into a number of clusters, where the points inside a cluster show a certain degree of closeness or similarity [1, 2]. It has been the subject of wide research arising in many application domains in engineering, business, medical, and social sciences. Clustering methods can be considered as either hard (crisp) [3, 4] or fuzzy [5, 6, 7] depending on whether a pattern belongs exclusively to a single cluster or to several clusters with different degrees. In hard clustering each point of the dataset belongs to exactly one cluster, a membership value of zero or one is assigned to each pattern, whereas in fuzzy clustering, a value between zero and one is assigned to each pattern by a membership function.

Clustering methods [3, 7] have been successfully used to segment an image into a number of clusters (segments). However, clustering-based segmentation techniques [8, 9] have used several control parameters, e.g., the predefined number of clusters to be found or some tunable thresholds. These parameters should be adjusted to obtain the best image segmentation. The choice of values for the various parameters is a nontrivial task. If quantitative evaluation function (known under the general term cluster validity method) of segmentation results is applied then a choice of values of parameters is simpler. Many criteria have been developed for determining cluster validity [10] [11], all of which have a common goal to find the clustering which results in compact clusters that are well separated.

The key questions in this paper are "Can the appropriate number of clusters be determined automatically? And if the answer is yes, how?" We survey several major existing solutions for cluster validity starting from the oldest fuzzy cluster validity functions [12], the partition coefficient (PC) and partition entropy (PE), that were first proposed in 1974, up to very recently reported ones, such as the partition coefficient and exponential separation (PCAES) index of Wu and Yang [24]. We implement all those methods for the popular fuzzy c-means (FCM) and k-means crisp clustering algorithms. There are broadly extended types of these two algorithms in the literature. Nevertheless, validity indexes are considered to be independent of clustering algorithms [25]. Thus, we will consider only the standard FCM and K-means clustering algorithms [6] for all validity indexes.

In addition, we propose two new validity indexes of our own. The first is based on Akaike's information criterion (AIC) [27]. Although AIC was extensively used in may other domains such as statistical estimation of model fitting, we employ it for the first time as a validation index, where the appropriate number of clusters can be determined automatically. The second index is developed from the well-established idea of cross-validation [29]. The performances of all the existing and proposed indexes in the domain of image segmentation are experimentally evaluated and compared on several test images under noisy conditions of varying degrees. As segmentation of medical images is of particular interest in our application of cluster analysis, a notable characteristic of our work here is the assessment of those indexes on 3D Magnetic Resonance (MR) datasets.

The rest of this paper is organized as follows: Section 2 presents the basics of the clustering algorithms. Several criteria to determine number of clusters are listed in Section 3. Experimental comparisons are presented in Section 4. Finally, Section 5 gives our conclusions.

2. STANDARD CLUSTERING ALGORITHMS

2.1. K-means algorithm

K-means clustering, also known as hard c-means clustering, is one of the simplest unsupervised classification algorithms. The procedure follows a simple way to classify the dataset through a certain number of clusters. The algorithm partitions a set of N vector X={x_j, $j=1,...,N$} into C classes c_i, $i=1, \ldots, C$, and finds a cluster centre for each class v_i denotes the centroid of cluster c_i such that an objective function of dissimilarity, for example a distance measure, is minimized. The objective function that should be minimized, when the Euclidean distance is selected as a dissimilarity measure, can be described as:

$$P = \sum_{i=1}^{C}\left(\sum_{k, x_k \in c_i}\|x_k - v_i\|^2\right), \quad (1)$$

where $\sum_{k, x_k \in c_i}\|x_k - v_i\|^2$ is the objective function within group i, and $\|x_k - v_i\|^2$ is a chosen distance measure between a data point x_k and the cluster centre v_i. The partitioned groups are typically defined by a ($C \times N$) binary membership matrix $U=(u_{ij})$, where the element u_{ij} is 1 if the j-th data point x_j belongs to group i, and 0 otherwise. That means:

$$u_{ij} = \begin{cases} 1 \ \ if \|x_j - v_i\|^2 \le \|x_j - v_k\|^2, \text{ for each } k \neq i \\ 0 \ \ otherwise \end{cases} \quad (2)$$

$$v_i = \frac{\sum_{x_j \in v_i} \sum_{j=1}^{N} x_j}{R_i} \quad (3)$$

where R_i is number of data point in class c_i.

2.2. Fuzzy C-means clustering method

Fuzzy c-means clustering (FCM) is a data clustering algorithm in which each data point belongs to a cluster with a degree specified by its membership grade. Bezdek [12] has proposed this algorithm as an alternative to earlier K-means clustering. FCM partitions a collection of N vector x_i, $i=1,\ldots,N$ into C fuzzy groups, and finds a cluster centre in each group such that an objective function of a dissimilarity measure is minimized. In FCM, the membership matrix U is allowed to have not only 0 and 1 but also the elements with any values between 0 and 1. This matrix satisfies the constrained:

$$\sum_{i=1}^{C} u_{ij} = 1, \forall j = 1, \ldots, N \quad (4)$$

The objective function of FCM can be formulated as follows:

$$P(u, c_1, \ldots, c_C) = \sum_{i=1}^{C}\sum_{j=1}^{N} u_{ij}^{m}\|x_j - v_i\|^2, \quad (5)$$

where u_{ij} is between 0 and 1; v_i is the cluster centre of fuzzy group i, and the parameter m is a weighting exponent on each fuzzy membership (in our implementation, we set it to 2). Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, updating of membership u_{ij} and the cluster centers v_j by:

$$v_i = \frac{\sum_{j=1}^{N} u_{ij}^{m} x_j}{\sum_{j=1}^{N} u_{ij}^{m}}. \quad (6)$$

$$u_{ij} = \frac{1}{\sum_{k=1}^{C}\left(\frac{\|x_j - v_i\|}{\|x_j - v_k\|}\right)^{2/(m-1)}} \quad (7)$$

3. CLUSTER VALIDITY INDEXES

In the literature, two categories of indexes can be found. The first category uses only the membership values, u_{ij}. The other one involves both the U matrix and the dataset itself. This section lists the indexes of both categories and proposes two new ones.

3.1 Category I: Indexes involving only the membership values

This category includes Bezdek's *the partition coefficient (PC)* [12], *the partition entropy coefficient (PE)* [13] and Dave's *modification of the PC index (MPC)* [14] to reduce its monotonic tendency.

3.2 Category II: Indexes involving the membership values and the dataset.

This category includes *the Xie-Beni index (S)* [16], the *Modified Xie-Beni index (XB)* [17], *the I index* [18], the *Davies-Bouldin index* (*DB*) [19], *the cluster validity measure (VM)* [20], the *Fukuyam-Sugeno index (FS)* [21], *the fuzzy hyper volume (FHV)* [22], *the average partition density (PA)* [19], the *partition density index (PD)* [19], and the *separation and compactness index (SC)* [23]. It also includes recently proposed indexes, such as the *partition coefficient and exponential separation (PCAES) index* [24], the *PBMF-index* [25] and t*he compose within and between scattering (CWB)* index [26].

3.3 Proposed indexes

3.3.1Index Based on Akaike's information criterion (AIC)

AIC technique was originally proposed by Akaike [27]. However, different schemes based on the *AIC* have been developed and used in different applications [28]. The classical *AIC* is defined as:

$$AIC = D_a + 2\mu\sigma^2, \quad (8)$$

where μ is the number of degrees of freedom of the model, and D_a can be equal to P from Eq.(1). The value of μ is given by $\mu(C) = (C-1)N + C$ in the case of soft clustering, while $\mu(C) = N + C$ in the case of hard clustering. The noise level σ can be estimated from

$$\sigma^2 = \frac{D_a(C^*)}{\rho N - \mu(C^*)} \quad (9)$$

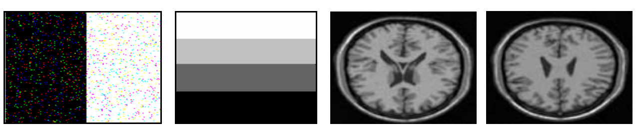
where C^* is the maximum number of cluster, ρ is the co-dimension of the model (ρ=1). The smaller the *AIC* value is, the better the clustering performance for the data set.

3.3.2 Index Based on cross-validation (V Index)

We now propose another index for cluster validity. This index is based on cross-validation, which is an old, standard tool in statistics [29]. The data are divided into two sets, one used for determining the clusters and the other one is used to validate the obtained clusters. The underlying idea here is to validate them on a dataset different from the one used for cluster estimation. For the task of image segmentation, the two subsets of data can be formed in several ways. One way, which we will follow, is to use an under-sampled version of the dataset for cluster estimation and the original dataset for validation. In our implementation, the under-sampled dataset is obtained by averaging every 2×2 pixels (2×2×2 voxels for 3D data) in the original dataset. One desirable effect here is that the resultant half-sized dataset contains smaller noise, which ought to lead to better cluster estimation.

4. EXPERIMENTAL RESULTS

The experiments were performed with several data sets. The first experiment consists of two simple synthetic images (synthetic1 and synthetic2), the first corrupted by 9% salt and pepper noise, and the other corrupted by Gaussian noise of standard deviation 20. The images are of size 142×145 pixels, as shown in Fig. 1(a), and Fig. 1(b), respectively. The second set includes simulated volumetric MR data consisting of ten classes, obtained from the classical simulated brain database of McGill University [30]. Two slices drawn from the simulated MR data is shown in Figs. 1(d) and 1(e).



(a) **(b)** **(c)** **(d)**

Fig.1: Test images: (a) Synthetic 1, (b) Synthetic 2, (c) and (d) two original slices from the 3D simulated data (slice 91 and slice 100).

Note that the true number of clusters for synthetic1, and synthetic2 is obviously 2 and 4 clusters, respectively. For the 3D simulated MR data, the correct number of clusters can be 8, 9 or 10 (as two clusters are indeed a combination of some of the remaining eight). The fuzzy c-means and k-means algorithms are applied independently on each dataset, and the aforementioned indexes are used to estimate the number of clusters. The outcome of each index on the different test images is reported in Table I. It can be concluded from Table I that in the case of soft partitioning only the *FHV* and I indexes yielded the true number of clusters on the 2D datasets, on which the *PBMF* index yielded almost the correct number as well. However, out of the three, only I and *PBMF* indexes gave the correct number on the 3D volumetric dataset. On the other hand, the proposed AIC and V indexes yielded correct results on the 3D volume but had erroneous ones on the 2D datasets.

On the other hand, in case of crisp partitioning, one can read from Table I that although many indexes give accurate results on the 3D volume, only the *I* and *PBMF* indexes yield correct or almost correct results on synthetic 2D images. The *AIC* index has poor performance, while the *V* index still provides accurate results on only the 3D MR data.

Table I. Number of clusters obtained by various indexes using the fuzzy c-means and k-means algorithms

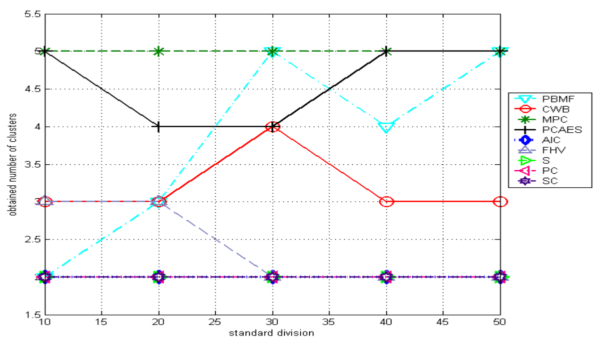
	Synthetic1		**Synthetic2**		**3D simulated data**	
	FCM	**K-means**	**FCM**	**K-means**	**FCM**	**K-means**
	Obtained number of clusters					
PD	3	2	2	2	10	2
PA	2	2	2	2	2	2
FHV	2	5	4	5	2	2
FS	5	4	2	7	6	10
PE	2	5	3	7	2	10
PC	5	5	3	7	2	10
S	2	2	2	2	2	2
XB	2	2	2	2	2	2
DB	2	4	7	5	10	10
I	2	3	4	4	10	2
VM	2	5	2	7	2	8
AIC	5	2	7	2	10	2
MPC	4	5	7	7	6	10
SC	2	2	2	2	2	2
PCAES	5	5	5	7	10	10
PBMF	3	2	5	4	10	7
CWB	3	4	3	6	3	4
V	5	5	6	7	9	10

Furthermore, the performance of each index against noise is evaluated when the two synthetic images are corrupted with zero-mean Gaussian noise with standard deviation ranging from 0 to 50. Note that the results in the previous table were obtained for Synthetic1 corrupted with 9% salt and pepper noise. Fig. 2 depicts the relationships between the number of clusters found by the several indexes and the noise standard deviation when the FCM algorithm is applied to the two synthetic images. It is clear that when *MPC*, *XB*, *AIC* and *SC* indexes are used, the found number of cluster is constant for various degrees of noise, while for the *CWB* index this relationship is unstable. As the noise level increases, the number of clusters obtained by the *FS* index increases. The *FHV, I*, *V*, and *PBMF* indexes show inconsistent behavior between the two images.

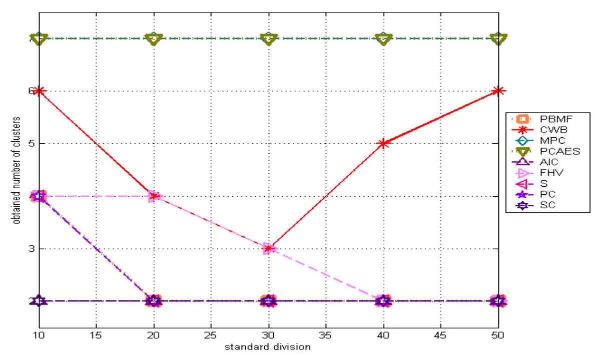
Analogously, on applying the k-means algorithm to the two noisy synthetic images, the corresponding relationships between the found number of clusters and noise standard deviation have revealed that the *MPC*, *S*, *PC*, *XB*, *PA*, *SC*, *V* and *AIC* indexes have constant outcomes for various levels of noise, whereas this relationship is unstable for the *CWB* and *PBMF* indexes.

5. CONCLUSIONS

In clustering, the role of a validity index is very important. Through the help of some indexes that measure the quality of clusters formed by an algorithm, it is hoped that the number of clusters that present an image may be determined automatically. We have surveyed 16 well-known such indexes and made a comprehensive comparison between these indexes for the task of image segmentation. We also proposed a new index based on Akaike's information criterion (*AIC*). In addition, a new index for the same task based on cross-validation has been proposed. All 18 indexes have been assessed on 2D and 3D data corrupted with noise of varying levels. To the best of our knowledge, no such comprehensive survey and comparison has been reported before in literature.



(a)



(b)

Fig.2: The relationships between the found number of clusters and noise standard deviation for *PBMF, CWB, MPC, SC,* PCAES, *AIC* , *FHV, S and PC* indexes when the fuzzy clustering is applied to (a) synthetic1 and (b) syntetic2 images.

From our experiments, one cannot conclude in general that a particular index would work well in all cases. But some cluster validity indexes may guide the selection of the appropriate number of clusters existing in a dataset. The *I* index has shown better results in that regard. However, the *MPC*, *XB*, *AIC* and *SC* indexes have demonstrated better, consistent performance on all test images under varying noise levels. More experiments, however, ought to be done to verify these findings. This, in addition to exploring other useful indexes, will be the focus of our future research.

6. REFERENCES

[1] S. Guha, et al., "CURE: An Efficient Clustering Algorithm for large Databases," In Proceedings of the ACM SIGMOD Conference, 1998.

[2] S. Theodoridis, and K. Koutroubas, Pattern Recognition, Academic Press, 1999.

[3] V.P. Guerrero-Bote, et al., "Comparison of neural models for document clustering," Int. Journal of Approximate Reasoning, vol. 34, pp.287-305, 2003.

[4] A. Likas, N. Vlassis, and J. Verbeek, "The global k-means clustering algorithm (Technical Report)," Computer Science Institute, University of Amsterdam, The Netherlands. ISA-UVA-01-02. 2001.

[5] G.L. Carl, "A fuzzy clustering and fuzzy merging algorithm, Technical Report," CS-UNR-101, 1999.

[6] J. Leski, "Towards a robust fuzzy clustering," *Fuzzy Sets and Systems*, vol. 137, no. 2, pp. 215-233, 2003.

[7] H. Arefi, M. Hahn, and F. Samadzadegan, "Comparison of clustering techniques applied to laser data," Geo-imagery Bridging continents xx-th ISPRS congress, Istanbul, Turkey, 2004.

[8] D. Pham, C. Xu, and J. Prince, "Current methods in medical image segmentation," Annu. Rev. Biomed. Eng. 2, 315-337, vol. 7(4), pp.201–206, 2000.

[9] T.N. Pappas, "An adaptive clustering algorithm for image segmentation," IEEE Trans. Signal process, vol. 40, pp.901-914, 1992.

[10] Y. Xu, G. Richard, and A. Brereton, "A comparative study of cluster validation indices applied to genotyping data," Chemometrics and Intelligent Laboratory Systems, vol. 78, pp. 30–40, 2005.

[11] K.L. Wu, and M.S. Yang, "A cluster validity index for fuzzy clustering," Pattern Recognition Letters, vol. 26, pp.1275–1291, 2005.

[12] J.C. Bezdek, "Pattern Recognition with Fuzzy Objective Function," Plenum, New York, 1981.

[13] J.C. Bezdeck, R. Ehrlich, and W. Full, "FCM: Fuzzy c-means algorithm," Computers and Geoscience, 1984.

[14] R.N. Dave, "Validating fuzzy partition obtained through c-shells clustering," Pattern Recognition Lett. 17, pp.613–623, 1996.

[15] M. Robubens, "Pattern classification problems and fuzzy sets," Fuzzy Sets Systems 1, 239–253, 1978.

[16] X.L. Xie, and G. Beni, "A validity measure for fuzzy clustering," IEEE Trans. PAMI, Vol.13, pp.841–847, 1991.

[17] G.L. Carl, "A fuzzy clustering and fuzzy merging algorithm, Technical Report," CS-UNR-101, 1999.

[18] U. Maulik, and S. Bandyopadhyay, "Performance evaluation of some clustering algorithms and validity indices," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, no. 12, 2002.

[19] D.L. Davies, and D.W. Bouldin, "A cluster separation measure," IEEE Trans. PAMI, vol. 1, pp. 224- 227, 1979.

[20] T. Liao, and BCompsc, "Image segmentation-Hybrid Method Combining Clustering and region merging," thesis at Monash University, 2003.

[21] Y. Fukuyama and M. Sugeno, "A new method of choosing the number of clusters for the fuzzy c-means method", Proceeding of fifth Fuzzy System Symp., pp. 247– 250, 1989.

[22] I. Gath, and A.B. Geva, "Unsupervised optimal fuzzy clustering," IEEE Trans. PAMI, vol.1, pp.773–781, 1989.

[23] N. Zahid, M. Limouri, and A. Essaid, "A new cluster-validity for fuzzy clustering," Pattern Recognition vol.32, pp.1089–1097, 1999.

[24] K.L Wu, and M.S. Yang, "A cluster validity index for fuzzy clustering," Pattern Recognition Letters, vol. 26,pp. 1275–1291, 2005.

[25] K.P. Malay, B. Sanghamitra, M. Ujjwal, "A study of some fuzzy cluster validity indices, genetic clustering and application to pixel classification", Fuzzy sets and systems vol. 155,pp.191-214, 2005.

[26]Yun Xu, Richard G. Brereton, "A comparative study of cluster validation indices applied to genotyping data", Chemometrics and Intelligent Laboratory Systems 78 (2005) 30–40

[27] H. Akaike, "On Entropy Maximization Principle, in *Applications of Statistics*," Krishnaiah, P.R., Ed., North Holland Publishing Company, pp. 27–41, 1977.

[28] J. Zhang, and J.M. Modestino, "A Model-fitting Approach to Cluster Validation with Application to Stochastic Model-based Image Segmentation," *IEEE Trans. PAMI,* vol. 12(10), pp. 1009–1017, 1990.

[29] Stone M., Cross-validatory choice and assessment of statistical predictions, Journal of the Royal Statistical Society, B36, pp. 111-133, 1974.

[30] BrainWeb: Simulated Brain Database. McConnell Brain Imaging Centre. Montreal Neurological Institute, McGill University, http://www.bic.mni.mcgill.ca/brainweb.