A New Clustering Validity Index for Fuzzy C Means Algorithm Based on Measure Of Disparity.

ADEKUNLE Y.A[1] , ALAO O.D[2] , EBIESUWA SEUN[3] , SARUMI JERRY[4] , AINAM JEAN-PAUL[5]

[1,2,3,5] Computer Science Department, Babcock University, Ilishan-Remo, Ogun State, Nigeria.

[4] Computer Science Department, Lagos State Polytechnic, Ikorodu, Lagos State, Nigeria.

ABSTRACT

Cluster validity indexes have been used to evaluate the fitness of partitions produced by clustering algorithms. This paper presents a new validity index for fuzzy clustering called inter-cluster and intra-cluster separation (IC2S) index. Therefore, we proposed the function of disparity which combines the intra and inter-cluster separation existing between the clusters. The results of comparative study show that the proposed IC2S index has high ability in producing a good cluster number estimate. This performance is achieved by taking into consideration the existing disparity between clusters. To assess the new validation index, two data sets (Fisher's IRIS and Butterfly data set) were used and the results show that IC2S outperforms other clustering validation index for fuzzy c-means.

Key words: cluster validity index, fuzzy clustering, and fuzzy c-means, fuzzy c-partitions.

INTRODUCTION

Fuzzy logic is a form of many-valued logic or probabilistic logic which deals with reasoning. The term "fuzzy logic" was first introduced in 1965 by A. Zadeh [1] as a new way to represent vagueness in everyday life. Fuzzy logic has been extended to handle the concept of partial truth, where the truth value may range between completely true and completely false [2]. Clustering [3, 4, 5, 6, 7] is an unsupervised classification method when the only data available are unlabeled, and no structural information about it is available. The objective of clustering is to find the data structure and also partition the data set into groups with similar individuals [8]. Amongst various fuzzy clustering algorithms, fuzzy c-means

(FCM) [9] is the basic one. Since it has some drawbacks, several algorithms have been developed to improve the performance. Although, several fuzzy cluster validity indices have been proposed to evaluate fuzzy clustering, they all suffer from the lack of combining the intra-cluster, inter-cluster and the geometry aspect of the cluster together. This paper focuses on developing a novel validation index for fuzzy clustering algorithm by taking in account the disparity among the clusters. This research is organized as follow. First, an overview on related works is presented (I), next, we briefly describe fuzzy c-means algorithms (II). In the third part of our paper, we present the proposed index by exhibiting the validation criteria and the FCM validation algorithm. Finally, two data sets were used to assess the proposed index and experimental results compared with a number of known validation indices discussed in [10, 11, 12, 13, 14].

I. RELATED WORKS

M. Ramze Rezaee et al., in [10] introduced a new validity index which assesses the average compactness and separation of fuzzy partitions generated by the fuzzy c-means algorithm. Using two data sets, they compared the performance of their index with a number of known validation indices [11, 12, 13, and 14]. The results of this study suggest that the new validation index can achieve the optimal result for any possible data set. By when the numerical representation chosen to describe the different object features do not properly discriminate between different classes, the validation index may fail. Also, the Euclidean norm (used in this model) may be unreliable for a specific data set. Finally, they applied the FCM by taking only a few samples of the weighting exponent (*$m = 2$)*, this parameter does not fit a careful analysis. V_{CWB} required very large values of *m*.

Maria Halkidi et al., in [15] review approaches and presented clustering validity checking approaches based on internal, external and relative criteria. They discussed the results of an experimental study based on widely known validity indices and finally illustrated the issues that are under-addressed by the recent approaches and proposed the research directions in the field. Though, no clustering validity index was suggested, our work is based on this research for they pointed out that quality measures that assess the quality of the partitioning need to be developed and i. intra-cluster quality, ii. inter-cluster separation and iii. geometry of the clusters need to be taken into account.

Chunhui Zhang et al., in [16] proposed a novel validity index for fuzzy-possibilistic c-means (FPCM) algorithm. It combines extended partition entropy and inter class similarity which is calculated from the fuzzy set point of view. The proposed index only requires the membership matrix and possibilistic (typicality) matrix, and is free from heavy distance computing. They finally compared their index with [10, 16] and results show its effectiveness. However, they did not consider similarity between the fuzzy sets and more other situations where the behavior of the proposed index may lead to worse result.

Yuangang Tang et al., in [17] proposed a new validation index for fuzzy clustering in order to eliminate the monotonically decreasing tendency as the number of clusters approaches to the number of data points and avoid the numerical instability of validation index when fuzzy weighting exponent increases. Two numerical examples have been presented to show the effectiveness of the proposed validation index. Yet, the proposed validation suffers from the lack of measures that assess the quality of the partitioning.

Weina Wang et al., in [18] introduced the fundamental concepts of cluster validity, and presented a review of fuzzy cluster validity indices available in the literature. They also conducted extensive comparisons of the mentioned indices in conjunction with the Fuzzy C-Means clustering algorithm on a number of widely used data sets, and made a simple analysis of the experimental results. No novel validation index is proposed on their work.

Moumen El-Melegy et al., in [19] sought 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 surveyed several key existing solutions for cluster validity in the domain of image segmentation and suggested two new indexes. Their novels indexes are only devoted to the domain of image recognition and therefore cannot be easily applied to other domains unless with adjustments.

There are many others publications, articles and journal on press concerning fuzzy clustering validation index [20, 12, 21, 22, and 23] among others. Most of these validity indices usually assume tacitly that data points having constant density to the clusters; however it is not sure of the real problems so far; there is no validation index for fuzzy c-means clustering algorithm which takes in account at the same time the three fundamental aspect of the clusters: geometry of the clusters, inter-cluster and intra-cluster separation.

II FUZZY C-MEAN ALGORITHM

Fuzzy c-mean (FCM) algorithm is an unsupervised clustering algorithm in which each data point belongs to a cluster with a degree specified by its membership grade. The description of the original algorithm dates back to 1973 [24, 25], derivatives have been described with modified definition for the norm and prototypes for the cluster centroids [26, 27, 28]. To find the centroid in each cluster and the grade of membership for each object in the clusters, FCM minimizes an objective function Jm, which is the weighted sum of squared errors within groups and is defined as follows:

$$J_m(U,V) = \sum_{j=1}^{n}\sum_{i=1}^{c} u_{ij}^{m} \left\|x_j - v_i\right\|^2 \quad (1)$$

$$1 < m < \infty$$

Where U is the membership matrix and is allowed to have not only 0 and 1 but also the elements with any values between 0 and 1. This matrix satisfies the constraint:

$$\sum_{i=1}^{c} u_{ij} = 1,\ \forall j = 1, \dots, N \qquad (2)$$

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 4, while most of preview papers set it to 2). v_i is defined by:

$$v_i = \frac{\sum_{j=1}^{n} u_{ij}^{m} x_j}{\sum_{j=1}^{n} u_{ij}^{m}} \qquad (3)$$

And u_{ij} (between 0 and 1) satisfies the constraint:

$$u_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{\left\|x_j - v_i\right\|}{\left\|x_j - v_k\right\|}\right)^{2/(m-1)}} \qquad (4)$$

II. NEW VALIDATION INDEX FOR FUZZY C-MEAN (IC2S)

a. Validation criteria

The FCM can find a partition of data for a fixed number of clusters known as objects. One objective of cluster validity is to determine automatically the optimal number of clusters [10]. There is a number of cluster validations available [10, 11, 16, 17, 18 and 27]. Some validity methods use only the membership values of a fuzzy partition of data. Among other functional, such indices are the partition coefficient $V_{PC,}$ the partition entropy V_{PE}, the proportion exponent and the uniform data functional. Table 2 lists a number of cluster validation indices, which are evaluated in our study.

a. Validation algorithm

$$\varepsilon = |u_{ik}(t+1) - u_{ik}(t)| \qquad (5)$$

Fuzzy c-mean algorithm:

```
Step 1: Choose a number of cluster which refers to a number of cluster to detect
Step 2: Randomly initializing the cluster center
Step 3: Repeat until the program has converged, that is where $\varepsilon \leq 0.001$
        Step 4: Computer the centroid of each cluster using formula eq. (7)
        Step 5: For each point, compute its coefficient of being in the cluster using formula eq. (6)
        Step 6: Generating new centroid for each cluster using formula eq. (6).
        Step 7: Computer $\varepsilon$ by using formula eq. (5), if $\varepsilon > 0.001$, go to step 3
End the loop.
End the algorithm with a collection of centroids (          ).
```

b. Proposed validation index

Although many validation index has been proposed, a reliable validation functional for the FCM must consider both inter and intra-cluster separation of a fuzzy c-partition. If only the inter-cluster separation is considered by the validation, the partition obtained considered each data as a separate cluster and neglect the intra-cluster separation; that is, the distance between each object of the cluster and the center c. Therefore a validation index which combines both criteria will have an optimal value for each partition. We have designed this validation index and called it Inter-Cluster and Intra-Cluster Separation (IC2S) index. IC2S is defined as followed:

$$\mathrm{VP}_{\mathrm{IC2S}}(\mathrm{U},\mathrm{V}) = \alpha\mathrm{Inter}(\mathrm{c}) + \beta\mathrm{Intra}(\mathrm{c}) \qquad (6)$$

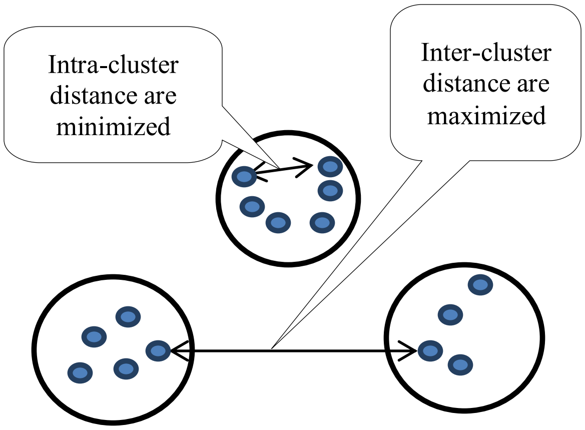
Intra-cluster distance are minimized
Inter-cluster distance are maximized

Figure 1: Intra and Inter-cluster representation

In order to get a reliable and functional validation index for FCM, the value of β in equation (6) should be as small as possible. When β tends to 0 (β → 0), the intra-cluster distance is also minimized. A classification is therefore achieved by choosing α such as α tends to the statistic variance of the population.

Centroid of a cluster c is determined by

$$c_i = \frac{1}{n}\sum_{i=1}^{k} u_{ik} \qquad (7)$$

The centroid is (typically) the mean of the points in the cluster.

The intra and inter-cluster separation are computed as followed:

For one cluster r:

$$Dr = \sum_i \sum_j \left\| x_i - x_j \right\|^2$$

(8)

Or

$$Dr = 2n_r \sum_i \| x_i - \bar{x} \|^2$$

For k cluster,

$$Wk = \sum_{r=1}^{k} \frac{1}{2n_r} Dr \qquad (9)$$

The inter-cluster distance is defined by:

```
If (number of cluster = 1) then

Else
```

and

$$\mathrm{intra(c)} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i=1}}^{\mathrm{n-1}} \|\mathrm{x_i} - \mathrm{x_{i+1}}\| \qquad (10)$$

Where n is the represents the number of object in the clusters. x_{i+1} And x_i represent two objects in the cluster at the ith iteration.

III. ANALYSIS AND RESULTS

Two data sets are used to assess our proposed index and the performance is compared with five well-known validation indices.

a. Data sets

The first data set used to assess our index is called Fisher's IRIS data. The full data set consists of 50 sets from each of three species of Iris (Iris Setosa, Iris Virginica and Iris versicolor). Four features were measured from each sample: le length and the width of the sepals and petals in centimeters. However, there's no need to pile up the sample numbers for our experiments. The first twenty sets of measurements for each species will suffice. These are reproduced in Table 1.

The second data set used to assess the index is called butterfly dataset from [33].

Table 1: First twenty specimens from each species included in Fisher (1936) Iris data

	Iris Setosa				**Iris versicolor**				**Iris Virginica**			
	Petal		Sepal		Petal		Sepal		Petal		Sepal	
	Length	Width	Length	Width	Length	Width	Length	Width	Length	Width	Length	Width
1	5.1	3.5	1.4	0.2	7.0	3.2	4.7	1.4	6.3	3.3	6.0	2.5
2	4.9	3.0	1.4	0.2	6.4	3.2	4.5	1.5	5.8	2.7	5.1	1.9
3	4.7	3.2	1.3	0.2	6.9	3.1	4.9	1.5	7.1	3.0	5.9	2.1
4	4.6	3.1	1.5	0.2	5.5	2.3	4.0	1.3	6.3	2.9	5.6	1.8
5	5.0	3.6	1.4	0.2	6.5	2.8	4.6	1.5	6.5	3.0	5.8	2.2
6	5.4	3.9	1.7	0.4	5.7	2.8	4.5	1.3	7.6	3.0	6.6	2.1
7	4.6	3.4	1.4	0.3	6.3	3.3	4.7	1.6	4.9	2.5	4.5	1.7
8	5.0	3.4	1.5	0.2	4.9	2.4	3.3	1.0	7.3	2.9	6.3	1.8
9	4.4	2.9	1.4	0.2	6.6	2.9	4.6	1.3	6.7	2.5	5.8	1.8
10	4.9	3.1	1.5	0.1	5.2	2.7	3.9	1.4	7.2	3.6	6.1	2.5
11	5.4	3.7	1.5	0.2	5.0	2.0	3.5	1.0	6.5	3.2	5.1	2.0
12	4.8	3.4	1.6	0.2	5.9	3.0	4.2	1.5	6.4	2.7	5.3	1.9
13	4.8	3.0	1.4	0.1	6.0	2.2	4.0	1.0	6.8	3.0	5.5	2.1
14	4.3	3.0	1.1	0.1	6.1	2.9	4.7	1.4	5.7	2.5	5.0	2.0
15	5.8	4.0	1.2	0.2	5.6	2.9	3.6	1.3	5.8	2.8	5.1	2.4
16	5.7	4.4	1.5	0.4	6.7	3.1	4.4	1.4	6.4	3.2	5.3	2.3
17	5.4	3.9	1.3	0.4	5.6	3.0	4.5	1.5	6.5	3.0	5.5	1.8
18	5.1	3.5	1.4	0.3	5.8	2.7	4.1	1.0	7.7	3.8	6.7	2.2
19	5.7	3.8	1.7	0.3	6.2	2.2	4.5	1.5	7.7	2.6	6.9	2.3
20	5.1	3.8	1.5	0.3	5.6	2.5	3.9	1.1	6.0	2.2	5.0	1.5
∑	100.7	69.6	28.7	4.7	119.5	55.2	85.1	26.5	131.2	58.4	113.1	40.9
Min.	4.3	2.9	1.1	0.1	4.9	2	3.3	1	4.9	2.2	4.5	1.5
Max	5.8	4.4	1.7	0.4	7	3.3	4.9	1.6	7.7	3.8	6.9	2.5
Mean	5.035	3.48	1.435	0.235	5.975	2.76	4.255	1.325	6.56	2.92	5.655	2.045
Media	5	3.45	1.4	0.2	5.95	2.85	4.45	1.4	6.5	2.95	5.55	2.05

n												
Variance	0.17	0.16	0.02	0.01	0.35	0.14	0.19	0.04	0.51	0.14	0.39	0.07
S.Dev.	0.42	0.40	0.14	0.09	0.59	0.37	0.44	0.19	0.71	0.38	0.62	0.27

Common validation index for fuzzy c-mean:
Table 2: Four validation functional for the fuzzy c-mean from [10]

Validation Index	Functional Description	Optimal cluster number
Partition coefficient	$V_{PC}(U) = \frac{1}{n}\left(\sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik}^2\right)$	Max ($V_{PC}(U, c_i, m)$)
Partition entropy	$V_{PE}(U) = \frac{1}{n}\left(\sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik}\log_a(u_{ik}$	Min $\left(V_{PE}(U, c_i, m)\right)$
Fukuyama and Sugeno	$V_{FS,m}(U, V; X) = \left(\sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik}^m\,(\lVert x_k - v_i \rVert^2 - \lVert v_i - \bar{v} \rVert_A^2)\right)$	Min $\left(V_{FS}(U, c_i, m)\right)$
Xie and Beni	$V_{XB}(U, V; X) = \frac{\sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^m \lVert x_k - v_i}{n\left(min\{v_i - v_j\}\right)}$	Min $\left(V_{XB}(U, c_i, m)\right)$
CWB, M. Ramze et al.	$V_{CWB}(U, V) = \propto Scat\,(c) + Dis(c)$	Min $\left(V_{CWB}(U, c_i, m\right.$
IC2S	$VP_{IC2S}\,(U, V) = \alpha Inter(c) + \beta Intra(c) \qquad (6)$	Min (VP_{IC2S} (U, c_i, m))

Where $Scat(c) = \frac{\frac{1}{c}\sum_{i=1}^{c}\lVert\sigma(v_i)\rVert}{\lVert\sigma(X)\rVert}$ and $Dis(c) = \frac{D_{max}}{D_{min}}\sum_{k=1}^{c}(\sum_{z}^{c}\lVert v_k - v_z \rVert)^{-1}$ [10]

CWB = Compose Within and between scattering

CONCLUSION AND FURTHER STUDIES

All through this research, we demonstrated the necessity of proposing a new clustering validation index for fuzzy c-means algorithm by taking into account the disparity between each object of the cluster. Next, by deriving a function of disparity and combining it with the new clustering validation index, we are sure that our model will converge faster to the value of so-fixed ε.

However, fuzzy c-means clustering and similar algorithms have problems with high dimensional data sets and a large number of prototypes [32], our proposed model can then suffer from this. More studies must be done in this area in order to overcome high dimensional data sets cases. Furthermore, instead of choosing β randomly as proposed in this model (β small implies a close intra-cluster separation), some can defined a mathematical formula enabling a more accurate computation of β.

REFERENCES

[1]. Zadeh, L.A (1965). “Fuzzy sets”, Information and Control 8 (3), pp. 338 – 353.

[2]. Novak, V., Perfilieva, I. and Mockor, J. (1999) Mathematical principles of fuzzy logic Dodrecht: Kluwer Academic. ISBN 0-7923-8595-0.

[3]. M.R. Anderberg, Cluster Analysis for Application, Academic Press, New York, 1973.

[4]. P.A. Devijver, J. Kittler, Pattern Recognition: A Statistical approach, Prentice-Hall, London, 1982.

[5]. J.A Hartigan, Clustering Algorithms, Wiley, New York, 1975

[6]. A.K. Jain, R.C. Dubes, Algorithms for Clustering Data, Prentice-Hall, Englewood Cliffs, NJ, 1988.

[7]. Y. G. Tang, F.C. Sun, Z.Q. Sun, Improved validation index for fuzzy clustering, in: American Control Conf., June 8 – 10, 2005, Portland, OR, USA.

[8]. Fuzzy clustering algorithms for mixed feature variables, Miin-Shen Yang, Pei-Yuang Hwang, De-hua Chen, February 2003.

[9]. Zadeh, L. A. et al. 1996 Fuzzy Sets, Fuzzy Logic, Fuzzy Systems, World Scientific Press, ISBN 981-02-2421-4.

[10]. A new cluster validity index for the fuzzy c-mean, M. Ramze Rezaee, B.P.F. Lelieveldt, J.H.C Reiber, 1997.

[11]. Xie, X.L., Beni, G.A., A validity measure for fuzzy clustering. IEEE Trans. Pattern Anal. Machine Intell. 13 (8), 841 – 847, Aug. 1991.

[12]. Bezdek, J.C., 1974. Cluster validity with fuzzy sets. J. Cybernet. 3 (3), 58 – 72.

[13]. Bezdek, J.C, 1975. Mathematical models for systematics and taxonomy. In: Estabrook, G. (Ed.), Proc. 8th Internat. Conf. Numerical Taxonomy. Freeman, San Francisco, CA, pp. 143 – 166.

[14]. Fukuyama, Y., Sugeno, M., 1989. A new method of choosing the number of clusters for fuzzy c-means method. In: Proc. 5th Fuzzy Syst. Symp., pp. 247 – 250.

[15]. Clustering Validity Checking Methods: Part II, Maria Halkidi, Yannis Batistakis, Michalis Vazirgiannis.

[16]. A Validity Index for Fuzzy and Possibilistic C-means Algorithms, Chunhui Zhang, Yiming Zhou, Trevor Martin.

[17] Improved Validation Index for fuzzy clustering, Yuangang Tang, Fuchun Sun,