

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/343641959>

A New Fuzzy Cluster Validity Index for Hyperellipsoid or Hyperspherical Shape Close Clusters With Distant Centroids

Article in IEEE Transactions on Fuzzy Systems · August 2020

DOI: 10.1109/TFUZZ.2020.3016339

CITATIONS

21

READS

114

2 authors:



Himanshu Mittal

Jaypee Institute of Information Technology

48 PUBLICATIONS 681 CITATIONS

[SEE PROFILE](#)



Mukesh Saraswat

Jaypee Institute of Information Technology

81 PUBLICATIONS 2,114 CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:



BIOGEOGRAPHY BASED OPTIMIZATION [View project](#)



Nature Inspired Algorithms [View project](#)

A new fuzzy cluster validity index for hyper-ellipsoid or hyper-spherical shape close clusters with distant centroids

Himanshu Mittal and Mukesh Saraswat

Abstract—Determining the correct number of clusters is essential for efficient clustering and cluster validity indices are widely used for the same. Generally, the effectiveness of a cluster validity index relies on two factors: (i) separation, defined by the distance between a pair of cluster centroids or a pair of data points belonging to different clusters and (ii) compactness which is determined in terms of the distance between a data point and a centroid or between a pair of data points belonging to the same cluster. However, the existing cluster validity indices for centroid-based clustering are unreliable when the clusters are too close, but corresponding centroids are distant. To mitigate this, a new cluster validity index, Saraswat-and-Mittal Index, has been proposed for hyper-ellipsoid or hyper-spherical shape close clusters with distant centroids, generated by fuzzy c-means. The proposed index computes compactness in terms of the distance between data points and corresponding centroids, while the distance between data points of disjoint clusters defines separation. These parameters benefit the proposed index in the analysis of close clusters with distinct centroids efficiently. The performance of the proposed index is validated against ten state-of-the-art cluster validity indices on artificial, UCI, and image datasets, clustered by the fuzzy c-means.

Index Terms—Cluster validity index, hyper-ellipsoid or hyper-spherical clusters, centroid-based clustering, fuzzy c-means.

I. INTRODUCTION

CLUSTERING, an unsupervised learning, explores the similarities and/or dissimilarities among the attributes of unlabelled data to form clusters. The data with homogeneous characteristics are clustered together while heterogeneous data are kept in disjoint clusters. The essence of a clustering method is to form coherent and contrast clusters efficiently, especially in data-driven applications like image processing, data mining, bioinformatics, social networking, and web-mining [1]. Generally, three aspects are involved in clustering, namely cluster tendency assessment, cluster analysis, and cluster validation [2]. The cluster tendency assessment evaluates the feasibility of clustering by exploring the cluster sub-structures. Cluster analysis explores such sub-structures to form the corresponding clusters. According to different clustering models, cluster analysis methods are categorized into centroid-based clustering, hierarchical clustering, distribution-based clustering, relational clustering, and density-based clustering. Finally,

cluster validation is the quantitative evaluation of identified clusters and termed as cluster validity index (CVI).

CVI resolves two fundamental issues of clustering [3], (i) finding the optimum number of clusters, available in a dataset and (ii) measuring how well the data fit into the formed clusters. Therefore, CVIs have been employed in many real-world clustering problems, such as image segmentation, post-genomic data analysis, text clustering, and transactional data [4]. Usually, CVIs analyze internal characteristics, such as homogeneity and heterogeneity, of the formed clusters to determine the optimal number of clusters. Here, homogeneity corresponds to the compactness of a cluster, while heterogeneity refers to the separation among clusters. This study considers such cluster attributes to propose a new CVI for the analysis of hyper-ellipsoid or hyper-spherical shape close clusters with distant centroids. The proposed CVI is primarily designed as a post-clustering measure for the centroid-based clustering results obtained from the fuzzy c-means (FCM) method.

Generally, centroid-based clustering is an iterative process to find K cluster centroids in a dataset by minimizing the total intra-cluster distance. K -means, FCM, and their variants [6]–[8] are some of the popular centroid-based clustering methods. In literature, many centroid-based CVIs exist. Ball and Hall [9] proposed the first CVI by averaging the distance of data points from their corresponding centroids. Partition Coefficient (PC), presented by Bezdek [6], is a fuzzy CVI which considers the fuzzy membership values of the formed clusters. Dunn [10] defined the Dunn index in terms of inter-cluster distance and intra-cluster distance among the data points. Moreover, the ratio of separation (distance between each centroid and the global centroid) and compactness (the distance of each point with its respective centroid) defines the Calinski-Harabasz Index (CHI) [11]. Further, Davies and Bouldin [12] introduced the Davies-Bouldin index (DBI), which measures the cluster similarity in terms of data density. Bezdek [13] considered the sum of squared membership values along with their logarithmic values to determine the Partition Entropy (PE) index. Fukuyama and Sugeno [14] used the variation between fuzzy compactness and separation to define Fukuyama and Sugeno Index (FSI). Furthermore, Pakhira et al. [15] proposed the Pakhira-Bandyopadhyay-Maulik-Fuzzy (PBMF) index for fuzzy clustering by analyzing intra-cluster compactness and inter-cluster separation. Wu and Yang [16] integrated the normalized partition coefficient with exponential separation to evaluate clustering and called it as the partition coefficient and

Himanshu Mittal, Department of Computer Science and Engineering, Jaypee Institute of Information Technology, Noida, India.
E-mail: himanshu.mittal224@gmail.com
Mukesh Saraswat, Department of Computer Science and Engineering, Jaypee Institute of Information Technology, Noida, India.

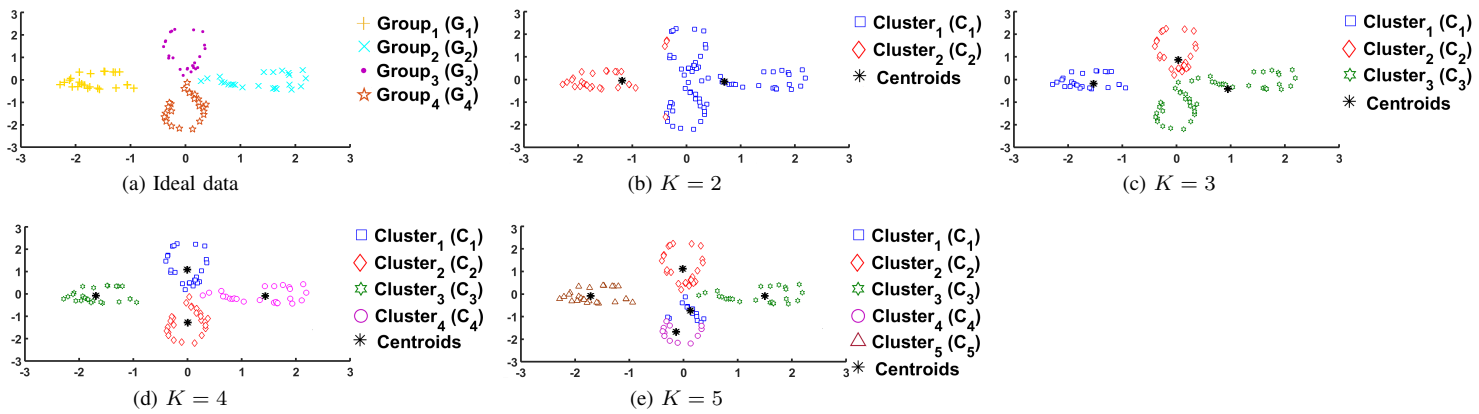
Fig. 1: Artificial petals data distribution [5] and corresponding clusters formed by FCM for different K values.

TABLE I: Existing CVIs.

S.No.	CVI	Formulation
1.	Dunn Index ($Dunn^+$)	$Dunn(K) = \min_{1 \leq s \leq K} \left(\min_{s+1 \leq t \leq K-1} \left(\frac{\min_{x_t \in C_s, x_j \in C_t} (x_i - x_j)}{\max_{1 \leq k \leq K} [\max_{x_t \in C_k} (x_i - x_j)]} \right) \right)$
2.	Partition coefficient index (PC^+)	$PC(K) = \frac{1}{N} \sum_{k=1}^K \sum_{i=1}^N \mu_{ik}^2$
3.	Calinski and Harabasz Index (CHI^+)	$CHI(K) = \frac{[\sum_{k=1}^K C_k v_k - \bar{v} ^2]}{K-1} / \frac{[\sum_{k=1}^K \sum_{x_i \in C_k} x_i - v_k ^2]}{N-K}$
4.	Davies and Bouldin Index (DBI^-)	$DBI(K) = \frac{1}{K} \sum_{k=1}^K \max_{j \neq k} \left\{ \frac{\left[\frac{1}{ C_j } \sum_{x_i \in C_j} x_i - v_j ^2 \right] + \left[\frac{1}{ C_k } \sum_{x_i \in C_k} x_i - v_k ^2 \right]}{ v_j - v_k ^2} \right\}$
5.	Partition entropy index (PE^-)	$PE(K) = \frac{1}{N} \sum_{k=1}^K \sum_{i=1}^N \mu_{ik} \log_2(\mu_{ik})$
6.	Fukuyama and Sugeno Index (FSI^-)	$FSI(K) = \sum_{k=1}^K \sum_{i=1}^N \mu_{ik}^m x_i - v_k ^2 - \sum_{k=1}^K \sum_{i=1}^N \mu_{ik}^m v_k - \bar{v} ^2$
7.	PBMF Index ($PBMF^+$)	$PBMF(K) = \frac{1}{K} \times \frac{\max_{j \neq k} \{ v_j - v_k \}}{\sum_{k=1}^K \sum_{i=1}^N \mu_{ik}^m x_i - v_k } \times \sum_{i=1}^N \mu_{i1} x_i - v_1 $
8.	PCAES Index ($PCAES^+$)	$PCAES(K) = \sum_{k=1}^K \sum_{i=1}^N \frac{\mu_{ik}^2}{\min_{1 \leq k \leq K} \{ \sum_{i=1}^N \mu_{ik}^2 \}} - \exp \left(\frac{-\min_{h \neq k} \{ v_k - v_h ^2 \}}{\left[\frac{1}{K} \sum_{k=1}^K v_k - \bar{v} ^2 \right]} \right)$
9.	Wu-and-Li Index (WLI^-)	$WLI(K) = \frac{\left[\sum_{k=1}^K \left(\frac{\sum_{i=1}^N \mu_{ik}^2 x_i - v_k ^2}{\sum_{i=1}^N \mu_{ik}} \right) \right]}{2 \times \left[\frac{1}{2} (\min_{i \neq j} \{ v_i - v_j ^2 \}) + \text{median}_{i \neq j} \{ v_i - v_j ^2 \} \right]}$
10.	V_R Index (V_R^-)	$V_R(K) = \sum_{k=1}^c \frac{(1/n_k) \sum_{i=1}^m x_i - v_k ^2 + (1/c) v_k - \bar{v} ^2}{(1/(c-1)) \sum_{j=1}^c v_j - v_k ^2}$

exponential separation (PCAES) index. Recently, Wu et al. [2] presented the Wu-and-Li index (WLI) to validate clusters with closely allocated centroids generated by FCM. Ren et al. [17] formulated a new validity index (V_R) with a new penalty function to analyze the fuzzy-based clusters.

Additionally, CVIs also consider other clustering models. For example, Conn_Index works on prototype-based clustering [18], while S-index uses hierarchical-based clustering [19]. Further, Sledge et al. [20] applied general CVIs on relational datasets. Moreover, clustering methods, like swarm intelligence-based clustering methods, use CVIs as their objective functions for optimal clustering [21]. A comprehensive survey of CVIs with comparison can be found in [4]. From the literature, it is discerned that no single validity index outperformed the others. Besides, more than one index can be used to obtain reliable results. Moreover, Pal and Bezdek [7] proved that the large value of the optimal cluster number is not good as it will generate a large number of clusters

with fewer data points. Hence, some CVIs use a constraint of the minimum centroid distance for mitigating this problem. Furthermore, the existing CVIs do not perform well on hyper-ellipsoid or hyper-spherical shape close clusters with distant centroids.

To elucidate the same, an artificial existing data distribution (Petals) [5] is considered and depicted in Fig. 1 with the corresponding FCM clustering results for $K = \{2, 3, 4, 5\}$. As illustrated in Fig. 1 (a), three groups (G_2, G_3 and G_4) are very close, while G_1 is well separated from others. Further, the clustering results for $K = 2$ and $K = 3$ are inappropriate, while $K = 5$ has reasonable cluster distribution. However, $K = 4$ represents the best clustering pattern. In addition, Table I tabulates the considered CVIs for the analysis of this data distribution. Fig. 2 reports the optimal cluster number suggested by the considered CVIs for $K = \{2, 3, \dots, 9, 10\}$. In the figure, the black dotted line illustrates the correct cluster number for this distribution. It is visible from Fig. 2 that no

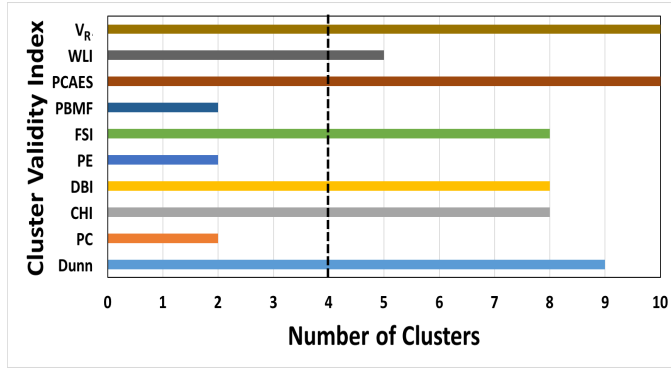


Fig. 2: Optimal cluster number suggested by different CVIs on the petals data distribution, depicted in Fig. 1 (a).

CVI has suggested the correct number of clusters for the given data distribution. However, only WLI suggested $K = 5$, which is close to the optimal number of clusters. All mentioned CVIs take distance of centroids to measure separation except Dunn. Therefore, other measures need to be considered for better analysis of separation in case of the distribution described above.

In order to efficiently analyze hyper-ellipsoid or hyper-spherical shape close clusters with distant centroids, this paper introduces a new CVI termed as Saraswat-and-Mittal index (SMI). The performance of SMI is validated on three datasets, namely artificial, UCI, and images. Experimental results show that the performance of SMI is comparatively stable than the considered CVIs and is accurate in determining the correct cluster numbers for FCM-based clustering.

The rest of the paper is organized as follows. Section II reviews FCM along with popular CVIs. The new CVI is presented in Section III. Section IV details the considered datasets and performance parameters which are used for the evaluation of the experiments. The experimental results are analyzed in Section V. Lastly, Section VI concludes the work.

II. PRELIMINARIES

In this section, the fuzzy c-means (FCM) is explained first, followed by a brief description of the existing CVIs for comparative analysis. The notations used in the descriptions are listed below.

- 1) X : Dataset having N data points,
 $X = \{x_1, x_2, \dots, x_N\}$
- 2) x_i : i^{th} data point
- 3) m : Level of cluster fuzziness
- 4) K : Number of clusters
- 5) C_k : k^{th} cluster
- 6) $|C_k|$: Total number of data points in k^{th} cluster
- 7) v_k : k^{th} cluster centroid
- 8) \bar{v} : Overall centroid of the dataset, defined as:
 $\bar{v} = \frac{1}{N} \sum_{i=1}^N x_i$
- 9) $\|x - y\|$: Standard Euclidian distance between x and y
- 10) μ_{ik} : Degree of membership of i^{th} data point with k^{th} cluster

A. Fuzzy C-Means (FCM)

FCM is a fuzzy distance-based clustering method that works on the principle of centroid-based clustering. In FCM, a data point is assigned to more than one cluster with a certain degree of membership. FCM was originally proposed by Dunn [10] and further improved by Bezdek [6]. It partitions the X iteratively to return a set of K fuzzy clusters and a partition matrix (U) to minimize the objective function defined in Eq. (1).

$$\sum_{i=1}^N \sum_{k=1}^K \mu_{ik}^m \|x_i - v_k\|^2, m \geq 1. \quad (1)$$

where, $\mu_{ik} \in [0, 1]$ defines the degree of membership of i^{th} data point with k^{th} cluster. m is the fuzzification parameter which controls the fuzziness among clusters. The valid range for m is $[1.5, 2.5]$ [7] however, researchers usually keep it as 2 [22]. The partition matrix (U) contains the value of each μ_{ik} as depicted in Eq. (2).

$$U = [\mu_{ik}] \in \mathbb{R}^{N \times K}, i = \{1, \dots, N\}, k = \{1, \dots, K\} \quad (2)$$

FCM optimizes Eq. (1) by updating μ_{ik} and v_k iteratively which are defined by Eqs. (3) and (4) respectively.

$$\mu_{ik} = \frac{1}{\sum_{j=1}^K \left(\frac{\|x_i - v_k\|}{\|x_i - v_j\|} \right)^{\frac{2}{m-1}}} \quad (3)$$

$$v_k = \frac{\sum_{i=1}^N \mu_{ik}^m x_i}{\sum_{i=1}^N \mu_{ik}^m} \quad (4)$$

Normally, the stopping criteria is defined as $\|U_{p+1} - U_p\| < \epsilon$, where p corresponds to the current iteration, while ϵ is a threshold, given by the user.

B. Existing Cluster Validity Index (CVI)

A cluster validity index purports to determine the correct clusters in X . The general procedure of a CVI is:

- Clustering method evolves k clusters for X , $\forall k = [1, K]$.
- The clustering results are evaluated by CVI, $\forall k = [1, K]$.
- The k with extreme (minimal or maximal) value of CVI defines the optimal partition for X .

For performance comparison, 10 state-of-the-art centroid-based CVIs are considered. The formulation of considered CVIs is summarized in Table I.

In the considered CVIs, the measurements of compactness and separation are the basic criteria to segregate the heterogeneous data points into disjoint clusters and homogeneous data points into the same cluster. Usually, the compactness is defined as the intra-cluster distance either between a pair of data points of a cluster, i.e., $\|x_i - x_j\|$, $i \neq j$, or between data points of a cluster and corresponding centroid i.e., $\|x_i - v_k\|$, where $x_i, x_j \in C_k$. The lower value of compactness corresponds to the high concentration of data points within a cluster. Moreover, the separation measures the inter-cluster distance either between a pair of centroids, as $\|v_k - v_h\|$, $k \neq h$, or between two heterogeneous data points from two different clusters, as $\|x_i - x_j\|$, where $x_i \in C_k$, $x_j \in C_h$. The large separation value signifies higher isolation among clusters. Dunn, CHI, and DBI use crisp distance measures, while PC, PE, FSI, PCAES, PBMF,

TABLE II: Result analysis of SMI on the Petals data distribution, depicted in Fig. 1 (a).

	$Co(K)$	$S(K)$	SMI^-
$K = 2$	0.9763	0.0799	12.2141
$K = 3$	1.2552	0.1552	8.0862
$K = 4$	0.8082	0.24	(3.3683)*
$K = 5$	0.8632	0.0518	16.7658
$K = 6$	0.8135	0.0447	18.2344
$K = 7$	0.8897	0.1385	6.4241
$K = 8$	1.1025	0.1536	7.1774
$K = 9$	0.8944	0.1663	5.3777
$K = 10$	0.9819	0.1698	5.7834

(*) : Optimal value of SMI^- .

WLI, and V_R consider fuzzy membership values for defining compactness and separation. Furthermore, CHI, FSI, and V_R measure separation as the distance between v_k and \bar{v} , while DBI, PCAES, PBMF, WLI, and V_R take the distance between v_h and v_k as separation value. However, Dunn index defines the separation in terms of the distance between data points belonging to distinct clusters, i.e. $\|x_i - x_j\|$. Generally, the extreme value of a CVI corresponds to the optimal K . PC, Dunn, CHI, PCAES, and PBMF prefer a maximum value for the optimal number of clusters (represented by a superscript ‘+’ sign). In contrast, the minimum values of PE, WLI, DBI, FSI, and V_R correspond to the optimal values (represented by a superscript ‘-’ sign).

III. PROPOSED CLUSTER VALIDITY INDEX

In this paper, a new CVI, Saraswat-and-Mittal index (SMI), is introduced to evaluate the clustering results of FCM. To measure the proximity of obtained clusters, SMI defines separation measure in terms of the distances among data-points of disjoint clusters rather than the distances among the centroids. On the contrary, SMI measures the compactness by considering the most expanded cluster. Since compactness should be small with a high separation value for the best fuzzy partition [23], the proposed SMI evaluates the ratio between the compactness and separation for K clusters as depicted in Eq. (5).

$$SMI(K) = \frac{Co(K)}{S(K)} \quad (5)$$

where, $Co(K)$ and $S(K)$ are fuzzy compactness and separation values for K clusters, calculated by Eqs. (6) & (7) respectively.

$$Co(K) = (K - 1) \left[\max_{\{1 \leq k \leq K\}} \left(\frac{\sum_{i=1}^N \mu_{ik}^2 \|x_i - v_k\|^2}{\sum_{i=1}^N \mu_{ik}} \right) \right] \quad (6)$$

$$S(K) = \min_{\{1 \leq s \leq K-1\}} \left(\min_{\{s+1 \leq t \leq K\}} \left(\text{dist}^2(C_s, C_t) \right) \right) \quad (7)$$

where,

$$\text{dist}(C_s, C_t) = \min_{\{x_i \in C_s, x_j \in C_t\}} (\|x_i - x_j\|) \quad (8)$$

As observed from Eq. (6), compactness computes the fuzzy intra-cluster distance in a K cluster system. It is the summation of the distances of each data point from the centroids with

μ membership degree. The maximum value among these distances corresponds to compactness in the proposed SMI. Further, Eq. (7) represents the separation as the inter-cluster variation in a K cluster system and considers the minimum distance among all the clusters based on data points, like the Dunn index. For better separation among clusters, a small value of $S(K)$ is desirable. Table II depicts the values returned by SMI for the petals data distribution, presented in Fig. 1 (a). From the table, it is noticeable that SMI reports the minimum value for $K = 4$, which corresponds to the optimal cluster number. Hence, SMI signifies attributes of an ideal CVI.

Further, the optimality of SMI to obtain a lower value is inevitable and mathematically proven by the theorem, given by Xie and Beni [24], as discussed below:

Theorem: If I is a validity index of a fuzzy partition of hard clusters and D_1 is the Dunn index for the corresponding hard partitions, then

$$I \leq \frac{1}{(D_1)^2} \quad (9)$$

It has already been proved by Dunn [10] that for hard clusters, the value of D_1 is always greater than one. Therefore, the above theorem can be redefined for a dataset with distinct substructures that if a fuzzy partition algorithm identifies the corresponding substructures, then the index value (I) must be less than one. On a similar concept, the validity of the proposed SMI index is proved in this paper.

Proof: Let $X = \{x_i | 1 \leq i \leq N\}$ is a dataset which has been optimally clustered by fuzzy partition into K clusters, having centroids $v_k, 1 \leq k \leq K$ and membership degrees $\mu_{ik}, 1 \leq i \leq N, 1 \leq k \leq K$. As defined earlier in Eq. (6), the optimal fuzzy compactness ($Co_{(opt)}$) of the clustered data is represented by Eq. (10).

$$Co_{(opt)} = (K - 1) \left[\max_{\{1 \leq k \leq K\}} \left(\frac{\sum_{i=1}^N \mu_{ik}^2 \|x_i - v_k\|^2}{\sum_{i=1}^N \mu_{ik}} \right) \right] \quad (10)$$

However, in case of hard K -partition, the corresponding compactness can be formulated as Eq. (11).

$$Co_{(h)} = (K - 1) \left[\max_{\{1 \leq k \leq K\}} \left(\frac{\sum_{x_i \in C_k} \|x_i - v_k\|^2}{|C_k|} \right) \right] \quad (11)$$

Ideally, it can be stated that,

$$Co_{(opt)} \leq Co_{(h)} \quad (12)$$

Let the centroid v_k lies within the boundary of cluster $C_k, 1 \leq k \leq K$ [24], i.e.,

$$\|x_i - v_k\|^2 \leq \text{dia}^2(C_k) \quad (13)$$

where, $\text{dia}(C_k)$ is defined by Eq. (14).

$$\text{dia}(C_k) = \max_{\{x_i, x_j \in C_k\}} \|x_i - x_j\|. \quad (14)$$

Thus, Eq. (12) can be rewritten as,

$$Co_{(opt)} \leq (K - 1) \left[\max_{\{1 \leq k \leq K\}} \left(\frac{\sum_{x_i \in C_k} \text{dia}^2(C_k)}{|C_k|} \right) \right] \quad (15)$$

which can further be simplified as Eq. (16).

$$Co_{(opt)} \leq (K - 1) \left[\max_{\{1 \leq k \leq K\}} \{\text{dia}^2(C_k)\} \right] \quad (16)$$

Further, as depicted in Eq. (7), the separation parameter (S) is represented by:

$$S = \min_{\{1 \leq s \leq K-1\}} \left(\min_{\{s+1 \leq t \leq K\}} \left(\text{dist}^2(C_s, C_t) \right) \right) \quad (17)$$

TABLE III: Details of considered artificial datasets (ADs).

Dataset	Number of instances	Number of features	True cluster numbers	Dataset	Number of instances	Number of features	True cluster numbers
AD1	4000	2	4	AD6	3000	2	20
AD2	3000	2	3	AD7	5000	2	15
AD3	100	2	4	AD8	5000	2	15
AD4	2000	2	4	AD9	1024	32	16
AD5	320	2	4	AD10-AD23	1351-10126	2-15	9

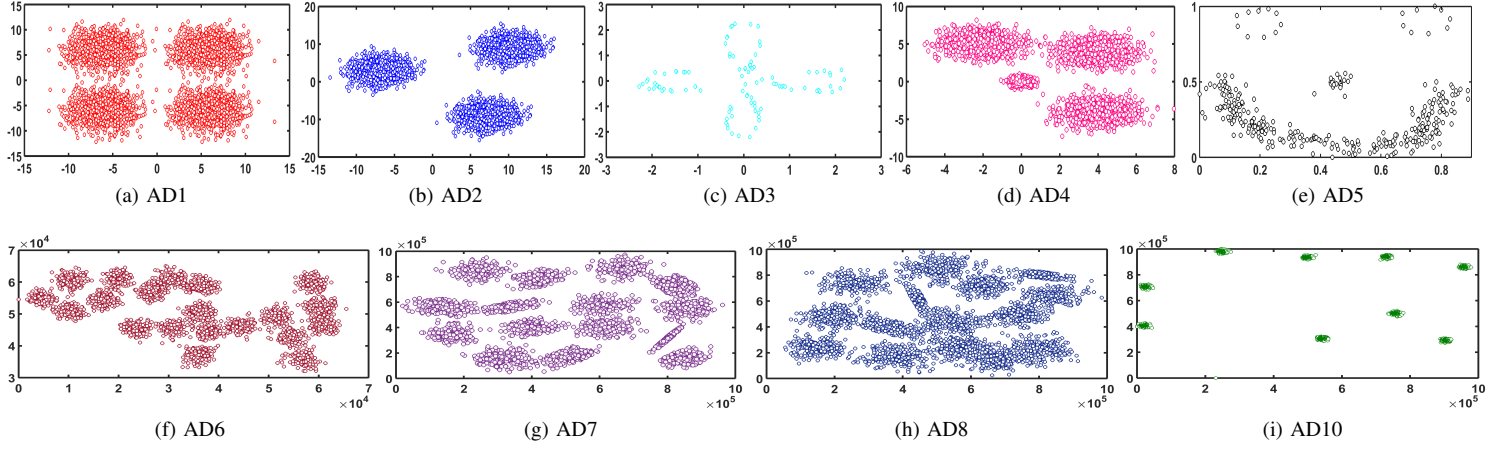


Fig. 3: Data distribution of artificial datasets (ADs).

From Eqs. (5), (16), and (17), SMI can be written as:

$$SMI \leq \frac{Co_{(opt)}}{S} \quad (18)$$

$$SMI \leq \frac{(K-1) * \max_{\{1 \leq k \leq K\}} dia^2(C_k)}{\min_{\{1 \leq s \leq K\}} (\min_{\{s+1 \leq t \leq K-1\}} (dist^2(C_s, C_t)))} \quad (19)$$

As per the equation of Dunn index, presented in Table I, Eq. (19) can be defined for any K as:

$$SMI \leq \frac{1}{(D_1)^2} \quad (20)$$

Hence, it is proved that SMI will attain arbitrarily low value if Dunn index increases without bound.

IV. EXPERIMENTAL SETUP

To evaluate the performance of the proposed SMI, experiments are simulated on a computer with Intel Core i3 of 2.8 GHz speed and 2 GB RAM using MATLAB R2015a. Since SMI is specifically designed for fuzzy clustering, FCM is used to perform the data clustering. However, FCM produces misleading results on large-scale datasets [25]. Therefore, experimental datasets are incorporated accordingly.

A. Datasets:

Three types of datasets, namely artificial datasets (AD1-AD23), UCI datasets (UCI1-UCI8), and image datasets (Img1 - Img5), have been considered for the performance evaluation of the SMI. The specification of all the ADs is detailed in Table III. Additionally, Fig. 3 depicts the data distribution of those artificial datasets which have the number of features as 2. AD1 and AD2 have been generated around four and three predefined cluster centroids with a normal distribution respectively, while AD3 to AD5 datasets are the standard synthetic datasets taken

from [26]. Further, AD6 consists of a large number of circular clusters, while AD7 and AD8 consist of Gaussian clusters with varying spatial complexity. On the contrary, AD9 corresponds to the Gaussian clusters of high-dimensional space. Similarly, AD10-AD23 are Gaussian clusters with varying dimensionality, ranging from 2 to 15, respectively. The artificial datasets (AD6-AD23) are taken from the publicly available clustering benchmark datasets [27]. The UCI datasets are considered from the UCI repository [28] and briefed in Table IV. The optimal partitioning for UCI datasets ranges from two to five. Additionally, the efficiency of the proposed SMI has been shown on the image segmentation problem by considering five images from the Berkeley Segmentation Dataset (BSDS 300) [29], as depicted in Fig. 4. Each image is a 64×64 RGB coloured image with a varying range of the optimal number of clusters based on the notable objects in the image as augmented by human experts.

B. Evaluation Criteria:

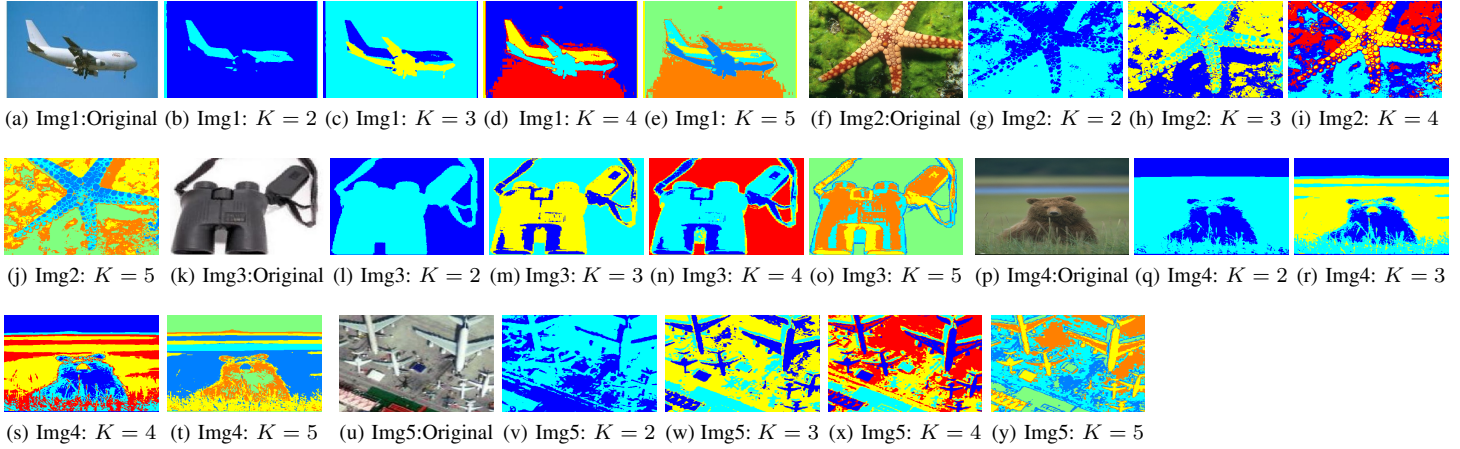
The interference of randomness in FCM is minimized by executing 50 rounds of FCM on each dataset with ϵ and m as 0.001 and 2 respectively. For each value of $K \in \{2, 3, \dots, k, \dots, 10\}$, K clusters formed by FCM are evaluated by all the CVIs in every round. The value of k , for which a CVI attains extreme value in a round, is designated as the optimal cluster number suggested by the respective CVI. This procedure is followed to identify the best cluster number (K_{best}) after all the rounds and equated as Eq. (21).

$$K_{best} = \arg \max_{2 \leq k \leq 10} O(k) \quad (21)$$

where, $O(k)$ represents the occurrence count of k for a CVI in 50 rounds.

TABLE IV: Details of UCI Datasets [28].

Abbr.	Dataset	Attribute types	No. of instances	No. of features	True cluster numbers
UCI1	Seeds	Real	210	7	3
UCI2	Breast Cancer Wisconsin (Original)	Integer	699	10	2
UCI3	Breast Cancer Wisconsin (Diagnostic)	Real	569	32	2
UCI4	Spectf Heart	Integer	267	44	2
UCI5	Connectionist Bench	Real	208	60	2
UCI6	Tarvel Review Ratings	Real	5456	25	5
UCI7	Online Shoppers Purchasing Intention	Integer, Real	12330	18	2
UCI8	Wholesale Customers	Integer	440	8	2

Fig. 4: Representative images (Img1-Img5) with clustering results on permissible K values (each colour denotes a cluster).

After all the rounds, the sensitivity of each CVI for a dataset is also analyzed. The formulation of sensitivity is defined by Eq. (22).

$$\text{Sensitivity (CVI)} = \frac{\text{count}(k = \#C)}{\text{Number of rounds}} \quad (22)$$

where, $\#C$ is the true cluster numbers for a dataset and $\text{count}(k = \#C)$ refers to the number of times k for a CVI matches $\#C$ in 50 rounds. Further, the value of $\#C$ for a dataset may be scalar (AD1 to AD23 and all UCI datasets except IRIS) or a range $[a, b]$, where $a \leq b$ (IRIS and image datasets).

Moreover, the effectiveness is computed to summarize the ability of a CVI in determining the correct cluster number over different categories of the considered datasets. Mathematically, the effectiveness of a CVI is formulated as Eq. (23).

$$E(\text{CVI}) = \frac{1}{M} \sum_{i=1}^M \delta(D_i), \quad (23)$$

where,

$$\delta(D_i) = \begin{cases} 1, & \text{if } K_{\text{best}} \in [a, b] \\ 0, & \text{else} \end{cases} \quad (24)$$

where, M is the total number of datasets in the considered category and D_i is the i^{th} dataset. The higher value of $E(\text{CVI})$ signify better performance of a CVI.

V. EXPERIMENTAL RESULTS

Tables V illustrates the results of the proposed and considered CVIs for artificial datasets. In Table V (a), the K_{best} value of SMI equals $\#C$ for all artificial datasets except AD5. As visible from Fig. 3 (e), the AD5 dataset has minimal data points in three clusters, and the fourth cluster is not spherical.

Thus, it leads to the inaccurate formation of clusters by FCM. The different formations of clusters on AD5 by FCM are depicted in Fig. 5 for $K = \{2, 3, 4, 5\}$. From the figure, it is observable that FCM is biased towards the high-density area in all K due to which no CVI determines the correct number of clusters for AD5. As far as other CVIs are concerned, all the considered CVIs (except PBMF and PCAES) performed well on four datasets, i.e. AD1, AD2, AD4, and AD23. However, for petals artificial dataset (AD3), only the proposed SMI gives accurate results with the sensitivity of 0.72 and rest CVIs do

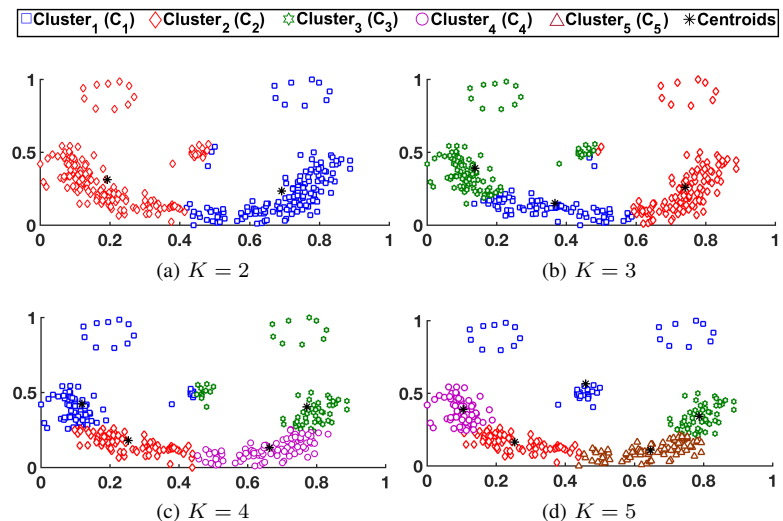
Fig. 5: Clustering results of FCM on AD5 for different K values.

TABLE V: Clustering results on artificial datasets in terms of (a) Cluster numbers decided by CVIs and (b) Sensitivity of each CVI.

(a)

Dataset	#C	$Dunn^+$	PC^+	CHI^+	DBI^-	PE^-	FSI^-	$PBMF^+$	$PCAES^+$	WLI^-	V_R^-	SMI^-
AD1	4	4	4	4	4	4	4	2*	8*	4	4	4
AD2	3	3	3	3	3	3	3	2*	9*	3	3	3
AD3	4	9*	2*	8*	8*	2*	8*	2*	10*	5*	10*	4
AD4	4	4	4	4	4	4	4	2*	10*	4	4	4
AD5	4	3*	2*	10*	2*	2*	9*	2*	10*	2*	10*	2*
AD6	20	20	2*	21*	27*	2*	22*	2*	59*	2*	2*	20
AD7	15	15	15	15	15	2*	15	2*	20*	15	20*	15
AD8	15	15	2*	15	15	2*	15	2*	20*	15	18*	15
AD9	16	16	2*	16	16	2*	20*	2*	20*	16	16	16
AD10	9	5*	9	9	9	9	9	2*	10*	9	9	9
AD11	9	8*	9	9	9	9	9	2*	10*	9	9	9
AD12	9	3*	9	9	9	9	9	2*	10*	9	9	9
AD13	9	2*	9	9	9	9	9	2*	10*	9	9	9
AD14	9	3*	9	9	9	9	9	2*	10*	9	9	9
AD15	9	5*	9	9	9	9	9	2*	10*	9	9	9
AD16	9	10*	9	9	9	9	9	2*	10*	9	9	9
AD17	9	5*	9	9	10*	9	9	2*	10*	9	9	9
AD18	9	5*	9	9	9	9	9	2*	10*	9	9	9
AD19	9	3*	9	9	9	9	9	2*	10*	9	9	9
AD20	9	6*	9	9	9	9	9	2*	10*	9	9	9
AD21	9	3*	9	9	9	9	9	2*	10*	9	9	9
AD22	9	2*	9	9	10*	9	9	3*	10*	9	9	9
AD23	9	9	9	9	9	9	9	2*	10*	9	9	9

* : incorrect or not within the acceptable range of clusters.

(b)

Dataset	$Dunn^+$	PC^+	CHI^+	DBI^-	PE^-	FSI^-	$PBMF^+$	$PCAES^+$	WLI^-	V_R^-	SMI^-
AD1	0.90	1.00	1.00	0.40	1.00	1.00	0.00	0.00	1.00	1.00	1.00
AD2	1.00	0.80	1.00	0.22	1.00	1.00	0.00	0.00	1.00	1.00	1.00
AD3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.72
AD4	0.38	1.00	1.00	0.24	1.00	1.00	0.00	0.00	1.00	0.96	1.00
AD5	0.16	0.00	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.10
AD6	0.78	0.00	0.16	0.00	0.00	0.22	0.00	0.00	0.16	0.00	0.84
AD7	0.72	0.68	0.68	0.20	0.00	0.68	0.00	0.00	0.68	0.00	0.75
AD8	0.80	0.00	0.78	0.18	0.00	0.78	0.00	0.00	0.78	0.00	0.86
AD9	0.58	0.00	0.64	0.72	0.00	0.00	0.00	0.00	0.67	0.69	0.82
AD10	0.10	0.98	0.96	0.50	0.98	0.96	0.00	0.00	0.96	0.96	0.98
AD11	0.00	0.96	0.96	0.64	0.96	0.96	0.00	0.00	0.96	0.96	0.96
AD12	0.00	0.98	0.98	0.56	0.98	0.98	0.00	0.00	0.98	0.78	0.98
AD13	0.02	1.00	0.98	0.76	1.00	1.00	0.00	0.00	0.98	0.98	1.00
AD14	0.00	1.00	1.00	0.66	1.00	1.00	0.00	0.00	1.00	1.00	1.00
AD15	0.22	1.00	1.00	0.70	1.00	1.00	0.00	0.00	1.00	0.96	1.00
AD16	0.00	0.78	0.78	0.60	1.00	0.78	0.00	0.00	0.78	1.00	1.00
AD17	0.00	0.62	0.62	0.38	1.00	0.62	0.00	0.00	0.62	0.62	1.00
AD18	0.00	0.94	0.86	0.54	0.94	0.86	0.00	0.00	0.86	0.90	1.00
AD19	0.34	1.00	0.88	0.64	1.00	0.88	0.00	0.00	0.88	0.82	1.00
AD20	0.00	0.98	0.90	0.64	0.98	0.98	0.00	0.00	0.90	0.92	0.98
AD21	0.00	0.96	0.96	0.64	0.98	0.96	0.00	0.00	0.96	0.88	0.96
AD22	0.00	1.00	0.94	0.46	1.00	1.00	0.00	0.00	0.94	0.94	1.00
AD23	0.82	1.00	0.94	0.52	1.00	1.00	0.00	0.00	0.94	0.94	1.00
AVG.	0.30	0.73	0.78	0.45	0.73	0.77	0.00	0.00	0.78	0.71	0.95

not suggest accurate cluster numbers even for a single time. It illustrates the efficiency of both compactness and separation components of SMI. Further, only Dunn and SMI determine the correct clusters for AD6. On AD7, PE, PBMF, PCAES, and V_R completely fail. Besides these CVIs, PC fails over AD8. On high-dimensional dataset (AD9), PC, PE, FSI, PBMF, and PCAES indicate inaccurate clusters. Moreover, Dunn, PBMF, and PCAES suggest incorrect cluster number for AD10-AD21. Furthermore, these CVIs, along with FSI, fail on AD17 and AD22.

The sensitivities of all the CVIs are tabulated in Table V (b). From the table, it is visible that the proposed SMI attains a sensitivity value of 1.00 on more than 50% of the

considered ADs. Additionally, it returns the sensitivity of more than 0.90 on AD10, AD11, AD12, AD20, and AD21. The overall sensitivity of SMI is 0.95 if AD5 is not considered, else it is 0.91. In both cases, sensitivity values are comparatively much better than all considered CVIs, which confirm that SMI generates stable results. Thus, the proposed SMI proves to be efficacious for all ADs.

The statistics for the UCI datasets are presented in Table VI. Table VI (a) validates that SMI returns correct cluster numbers for all the UCI datasets. Furthermore, the proposed SMI achieves the highest sensitivity, i.e. 1.00, for all UCIs except UCI6 for which it returns 0.68. For the UCI1 dataset, only Dunn, FSI, and SMI give competitive results. Furthermore,

TABLE VI: Clustering results on UCI datasets in terms of (a) Cluster numbers decided by CVIs and (b) Sensitivity of each CVI.

(a)

Dataset	#C	$Dunn^+$	PC^+	CHI^+	DBI^-	PE^-	FSI^-	$PBMF^+$	$PCAES^+$	WLI^-	V_R^-	SMI^-
UCI1	3	3	2*	2*	2*	2*	3	2*	10*	2*	2*	3
UCI2	2	3*	2	2	2	2	3*	2	9*	2	2	2
UCI3	2	2	2	10*	2	2	5*	2	10*	2	6*	2
UCI4	2	2	2	2	4*	2	10*	2	10*	2	10*	2
UCI5	2	5*	2	2	2	2	10*	2	10*	2	10*	2
UCI6	5	2*	2*	8*	9*	2*	10*	2*	10*	2*	10*	5
UCI7	2	2	2	10*	2	2	10*	2	10*	2	10*	2
UCI8	2	5*	2	3*	5*	2	10*	2	10*	3*	10*	2

* : incorrect or not within the acceptable range of clusters.

(b)

Dataset	$Dunn^+$	PC^+	CHI^+	DBI^-	PE^-	FSI^-	$PBMF^+$	$PCAES^+$	WLI^-	V_R^-	SMI^-
UCI1	0.98	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	1.00
UCI2	0.00	1.00	1.00	1.00	1.00	0.00	1.00	0.00	1.00	1.00	1.00
UCI3	1.00	1.00	0.00	0.60	1.00	0.00	0.80	0.00	1.00	0.02	1.00
UCI4	1.00	1.00	1.00	0.02	1.00	0.00	1.00	0.00	1.00	0.00	1.00
UCI5	0.00	1.00	1.00	0.66	1.00	0.00	1.00	0.00	1.00	0.00	1.00
UCI6	0.80	0.00	0.45	0.40	0.00	0.00	0.00	0.00	0.00	0.32	0.68
UCI7	1.00	1.00	0.00	0.70	1.00	0.00	0.50	0.00	1.00	0.00	1.00
UCI8	0.00	1.00	0.00	0.04	1.00	0.00	0.78	0.00	0.00	0.00	1.00
AVG.	0.60	0.75	0.44	0.43	0.75	0.13	0.64	0.00	0.63	0.17	0.96

TABLE VII: Clustering results on image datasets in terms of (a) Cluster numbers decided by CVIs and (b) Sensitivity of each CVI.

(a)

Dataset	#C	$Dunn^+$	PC^+	CHI^+	DBI^-	PE^-	FSI^-	$PBMF^+$	$PCAES^+$	WLI^-	V_R^-	SMI^-
Img1	[2, 3]	5*	2	9*	2	2	7*	2	10*	2	4*	2
Img2	[2, 4]	10*	2	3	2	2	5*	2	10*	2	4	2
Img3	[2, 3]	8*	2	10*	2	2	4*	2	10*	2	2	2
Img4	[3, 5]	10*	2*	6*	2*	2*	6*	2*	10*	2*	2*	3
Img5	[3, 4]	4	2*	9*	2*	2*	6*	2*	10*	2*	6*	4

* : incorrect or not within the acceptable range of clusters.

(b)

Dataset	$Dunn^+$	PC^+	CHI^+	DBI^-	PE^-	FSI^-	$PBMF^+$	$PCAES^+$	WLI^-	V_R^-	SMI^-
Img1	0.00	1.00	0.00	0.82	1.00	0.00	1.00	0.00	1.00	0.46	1.00
Img2	0.00	1.00	1.00	0.90	1.00	0.00	1.00	0.00	1.00	1.00	1.00
Img3	0.00	1.00	0.00	1.00	1.00	1.00	1.00	0.00	1.00	0.46	1.00
Img4	0.00	0.00	0.00	0.00	0.00	0.16	0.32	0.00	0.00	0.00	1.00
Img5	0.98	0.00	0.00	0.12	0.00	0.00	0.00	0.00	0.00	0.00	1.00
AVG.	0.20	0.60	0.20	0.57	0.60	0.23	0.66	0.00	0.60	0.38	1.00

PCAES does not return the correct cluster number for any UCI dataset. Moreover, FSI fails on every UCI dataset except UCI1, while V_R suggests the correct clusters for UCI2 only. Therefore, results on UCI datasets inferred that SMI can efficiently analyze the data points and performs relatively stable.

Further, the clustering results of the FCM method on the image dataset for $K \in \{2, 3, \dots, 5\}$ are illustrated in Fig. 4. Table VII represents the quantitative analysis of the results returned by proposed SMI and other CVIs on image dataset. In the table, SMI gives correct cluster numbers for all the images. Likewise, each object is distinct in the images corresponding to the suggested cluster numbers by SMI, as depicted in Fig. 4. For example, the important objects in Img1, Img2, and Img3 are plane, fish, and binoculars, respectively, which are easily recognizable in clustering results for $K = 2$. Thus, $K = 2$ is a reasonable cluster number for the corresponding images,

returned by SMI. Further, the bear is the main object in Img4, which is easily visualizable for $K = 3$ [see Fig. 4 (l)] and is determined by SMI only. Although the bear is recognizable in $K = 4$ and $K = 5$, there are many noisy spots. For Img5, SMI returns $K = 4$ in which the important objects, i.e. aircrafts, are distinctly segmented. Moreover, Table VII (b) illustrates the average sensitivity of the considered CVIs. SMI attains the average sensitivity of 1.0, followed by PBMF index, which is 0.66. Thus, the proposed CVI is favorable for predicting the optimal number of clusters in an image.

Table VIII summarizes the experimental results by calculating the effectiveness of all CVIs for artificial, UCI, and image datasets. For artificial datasets, the proposed CVI attains the maximum effectiveness, i.e. 0.95. Likewise, for both UCI and image datasets, only SMI has 1.0 effectiveness. The average effectiveness of SMI is 0.98, which is the best among all the considered indices.

TABLE VIII: Effectiveness of all CVIs

Dataset	$Dunn^+$	PC^+	CHI^+	DBI^-	PE^-	FSI^-	$PBMF^+$	$PCAES^+$	WLI^-	V_R^-	SMI^-
AD	0.34	0.78	0.86	0.78	0.73	0.82	0.00	0.00	0.86	0.78	0.95
UCI	0.50	0.75	0.37	0.50	0.75	0.12	0.75	0.00	0.62	0.12	1.00
IMG	0.20	0.60	0.20	0.60	0.60	0.00	0.60	0.00	0.60	0.40	1.00
AVG.	0.35	0.71	0.48	0.63	0.69	0.31	0.45	0.00	0.69	0.43	0.98

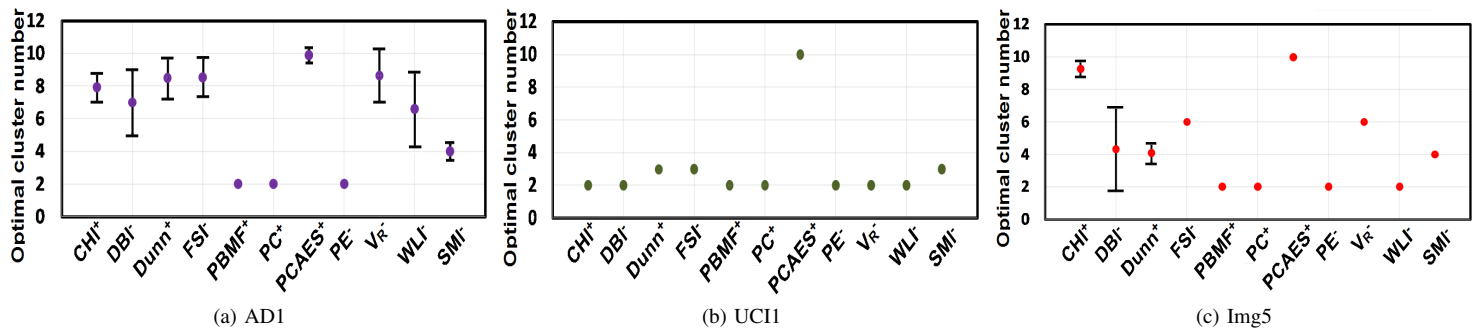


Fig. 6: Error-bar plots with standard deviation of considered CVIs on representative dataset of (a) artificial dataset, (b) UCI dataset, and (c) image dataset

Furthermore, Fig. 6 illustrates the error-bar plots with the standard deviation of considered CVIs on representative artificial (AD1), UCI (UCI1), and image (Img5) datasets. Evidently, SMI attains a quite low variation (or error) around the suggested optimal cluster values on every dataset. Therefore, SMI is comparatively more consistent than other CVIs. Consequently, it is pertinent to state that the proposed SMI is more efficient and stable than the state-of-the-art CVIs for centroid-based clustering.

VI. CONCLUSION

This paper introduces a new cluster validity index, Saraswat-and-Mittal Index (SMI), for hyper-ellipsoid or hyper-spherical shape close clusters with distant centroids, obtained from FCM. The proposed SMI is defined as the ratio of compactness of each cluster and separation among all the clusters. SMI measures separation in terms of the distance between data points of disjoint clusters. This consideration is favourable for handling the closely allocated clusters with distant centroids. The experimental evaluation manifested that SMI has performed better and stable than the 10 state-of-the-art considered CVIs on artificial, UCI, and image datasets. Some worthwhile aspects of evaluations are as follows.

- SMI is efficient in reporting optimal clusters for all artificial datasets except AD5. However, none of the CVI reported correct cluster numbers for AD5 due to the presence of a very high dense and non-spherical cluster as compared to the other three sparse clusters. It also affects the clustering results of the FCM method. Correspondingly, it is reasonable to consider the splitting of AD5 into two clusters as done by proposed SMI.
- The considered UCI datasets are the standard clustering datasets with the varied number of features on which SMI outperformed all the considered CVIs. Further, it is also interesting to note the stability of the results, produced by

SMI, which signifies that the cluster number suggested by SMI is either optimum or near to optimum in each round.

- The performance of SMI is further evaluated on pixel intensities of RGB coloured images, having a range of optimal clusters as augmented by domain experts. Although the shape of a multi-dimensional dataset is infeasible to determine and state whether the distribution is hyper-ellipsoid or hyper-spherical shape close clusters, clustering is an efficient method for image segmentation. On a similar application, SMI has suggested the correct cluster numbers within the allowable range for all the images. As the clustering quality of an image is subjective to application purpose; therefore, it is pertinent to state that the proposed SMI is beneficial in image clustering.

Henceforth, it is evident that the proposed SMI is stable and efficient. It may be used as an alternative for discovering optimal cluster numbers in centroid-based clustering problems, especially FCM. In the future, some research directions that will explore the dimensions of proposed SMI are as follows:

- The performance of SMI on noisy datasets and clustering applications is yet to be witnessed. Moreover, clustering methods other than centroid-based clustering, like hierarchical clustering, relational clustering, distribution-based clustering, and density-based clustering [30] may cluster the datasets differently, and SMI may fail on such clusters. However, with slight changes, the proposed CVI may be used for analyzing such clustering results.
- As SMI is a fuzzy CVI, it needs the membership matrix for evaluation purposes. However, a new variant of SMI may be designed for hard clustering by modifying the membership matrix parameter.
- As FCM does not correctly cluster the datasets having a very large number of clusters or features, SMI would not perform well on such datasets [25]. However, its performance may be studied by considering more robust fuzzy centroid-based clustering methods.

- Clustering is an aid to image analysis. This study analyzed RGB images only. Therefore, the performance of SMI may also be evaluated on other image features like pixel position or texture.
- The computational complexity of the compactness parameter is $\mathcal{O}(n)$ while the separation parameter has $\mathcal{O}(n^2)$, where n corresponds to the number of features. Although such computational complexity is problematic on the large value of n , it requires high-performance hardware. It also presents an open research area for exploring the various computational procedures for the same.

ACKNOWLEDGMENT

Authors are thankful to the Science and Engineering Research Board, Department of Science & Technology, Government of India, New Delhi for funding this work as part of the project (ECR/2016/000844).

REFERENCES

- [1] J.-Y. Jiang, R.-J. Liou, and S.-J. Lee, "A fuzzy self-constructing feature clustering algorithm for text classification," *IEEE transactions on knowledge and data engineering*, vol. 23, pp. 335–349, 2011.
- [2] C.-H. Wu, C.-S. Ouyang, L.-W. Chen, and L.-W. Lu, "A new fuzzy clustering validity index with a median factor for centroid-based clustering," *IEEE Transactions on Fuzzy Systems*, vol. 23, pp. 701–718, 2015.
- [3] M. K. Pakhira, S. Bandyopadhyay, and U. Maulik, "Validity index for crisp and fuzzy clusters," *Pattern recognition*, vol. 37, pp. 487–501, 2004.
- [4] E. Hancer and D. Karaboga, "A comprehensive survey of traditional, merge-split and evolutionary approaches proposed for determination of cluster number," *Swarm and Evolutionary Computation*, vol. 32, pp. 49–67, 2017.
- [5] C.-H. Li, B.-C. Kuo, and C.-T. Lin, "Lda-based clustering algorithm and its application to an unsupervised feature extraction," *IEEE Transactions on Fuzzy Systems*, vol. 19, no. 1, pp. 152–163, 2011.
- [6] J. C. Bezdek, "Cluster validity with fuzzy sets," *Journal of Cybernetics*, vol. 3, pp. 58–73, 1973.
- [7] N. R. Pal and J. C. Bezdek, "On cluster validity for the fuzzy c-means model," *IEEE Transactions on Fuzzy systems*, vol. 3, pp. 370–379, 1995.
- [8] M.-C. Chiang, C.-W. Tsai, and C.-S. Yang, "A time-efficient pattern reduction algorithm for k-means clustering," *Information Sciences*, vol. 181, pp. 716–731, 2011.
- [9] G. H. Ball and D. J. Hall, "Isodata, a novel method of data analysis and pattern classification," Stanford research institute, Menlo Park CA, Tech. Rep., 1965.
- [10] J. C. Dunn, "A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters," *Journal of Cybernetics*, vol. 3, pp. 32–57, 1973.
- [11] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics-theory and Methods*, vol. 3, pp. 1–27, 1974.
- [12] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE transactions on pattern analysis and machine intelligence*, vol. 1, pp. 224–227, 1979.
- [13] J. C. Bezdek, "Objective function clustering," in *Pattern recognition with fuzzy objective function algorithms*. Springer, 1981, pp. 43–93.
- [14] Y. Fukuyama and M. Sugeno, "A new method of choosing the number of clusters for the fuzzy c-mean method," in *Proc. of Fuzzy System Symposium*, 1989, pp. 247–250, Japan.
- [15] M. K. Pakhira, S. Bandyopadhyay, and U. Maulik, "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.
- [16] K.-L. Wu and M.-S. Yang, "A cluster validity index for fuzzy clustering," *Pattern Recognition Letters*, vol. 26, pp. 1275–1291, 2005.
- [17] M. Ren, P. Liu, Z. Wang, and J. Yi, "A self-adaptive fuzzy c-means algorithm for determining the optimal number of clusters," *Computational intelligence and neuroscience*, vol. 2016, pp. 1–12, 2016.
- [18] K. Tasdemir and E. Merényi, "A validity index for prototype-based clustering of data sets with complex cluster structures," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 41, pp. 1039–1053, 2011.
- [19] Z. Liang, P. Zhang, and J. Zhao, "Optimization of the number of clusters in fuzzy clustering," in *Proc. of International Conference on Computer Design and Applications*, 2010, pp. 580–584, China.
- [20] I. J. Sledge, J. C. Bezdek, T. C. Havens, and J. M. Keller, "Relational generalizations of cluster validity indices," *IEEE Transactions on Fuzzy Systems*, vol. 18, pp. 771–786, 2010.
- [21] A. Jose-Garcia and W. Gómez-Flores, "Automatic clustering using nature-inspired metaheuristics: a survey," *Applied Soft Computing*, vol. 41, pp. 192–213, 2016.
- [22] M. Huang, Z. Xia, H. Wang, Q. Zeng, and Q. Wang, "The range of the value for the fuzzifier of the fuzzy c-means algorithm," *Pattern Recognition Letters*, vol. 33, pp. 2280–2284, 2012.
- [23] N. Bharill and A. Tiwari, "Enhanced cluster validity index for the evaluation of optimal number of clusters for fuzzy c-means algorithm," in *Proc. of International Conference on Fuzzy Systems*. IEEE, 2014, pp. 1526–1533, China.
- [24] X. L. Xie and G. Beni, "A validity measure for fuzzy clustering," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 13, pp. 841–847, 1991.
- [25] L. Dalton, V. Ballarin, and M. Brun, "Clustering algorithms: on learning, validation, performance, and applications to genomics," *Current genomics*, vol. 10, pp. 430–445, 2009.
- [26] "6 functions for generating artificial datasets - file exchange - matlab central," <https://in.mathworks.com/matlabcentral/fileexchange/41459-6-functions-for-generating-artificial-datasets>, (Accessed on 01/23/2018).
- [27] P. Fränti and S. Sieranoja, "K-means properties on six clustering benchmark datasets," pp. 4743–4759, 2018. [Online]. Available: <http://cs.uef.fi/sipu/datasets/>
- [28] "Uci machine learning repository: Data sets," <http://archive.ics.uci.edu/ml/datasets.html>, (Accessed on 01/23/2018).
- [29] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. of International Conference on Computer Vision*, 2001, pp. 416–423, Canada.
- [30] M. R. Anderberg, *Cluster analysis for applications: probability and mathematical statistics: a series of monographs and textbooks*, 1st ed. Academic press, 2014.