A New Fuzzy Cluster Validity Index for Hyperellipsoid or Hyperspherical 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: first, separation, defined by the distance between a pair of cluster centroids or a pair of data points belonging to different clusters and second, 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 in this article for hyperellipsoid or hyperspherical 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, whereas 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*—Centroid-based clustering, cluster validity index (CVI), fuzzy c-means (FCM), hyperellipsoid or hyperspherical clusters.**

I. Introduction

CLUSTERING, an unsupervised learning, explores the similarities and/or dissimilarities among the attributes of unlabeled 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, such as 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 substructures. Cluster analysis explores such substructures 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], first, finding the optimum number of clusters available in a dataset and second, 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, postgenomic 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, whereas heterogeneity refers to the separation among clusters. This article considers such cluster attributes to propose a new CVI for the analysis of hyperellipsoid or hyperspherical shape close clusters with distant centroids. The proposed CVI is primarily designed as a postclustering 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 intracluster distance. K-means, FCM, and their variants [6]–[8] are some of the popular centroid-based clustering methods. In the 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 that considers the fuzzy membership values of the formed clusters. Dunn [10] defined the Dunn index in terms of intercluster distance and intracluster 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]. Furthermore, 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 intracluster

Manuscript received April 7, 2020; revised June 28, 2020; accepted August 7, 2020. Date of publication August 13, 2020; date of current version November 1, 2021. This work was supported by the Science and Engineering Research Board, Department of Science and Technology, Government of India, New Delhi, under Project ECR/2016/000844. *(Corresponding author: Himanshu Mittal.)*

The authors are with the Department of Computer Science and Engineering, Jaypee Institute of Information Technology, Noida 201309, India (e-mail: himanshu.mittal224@gmail.com; saraswatmukesh@gmail.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TFUZZ.2020.3016339.

Digital Object Identifier 10.1109/TFUZZ.2020.3016339

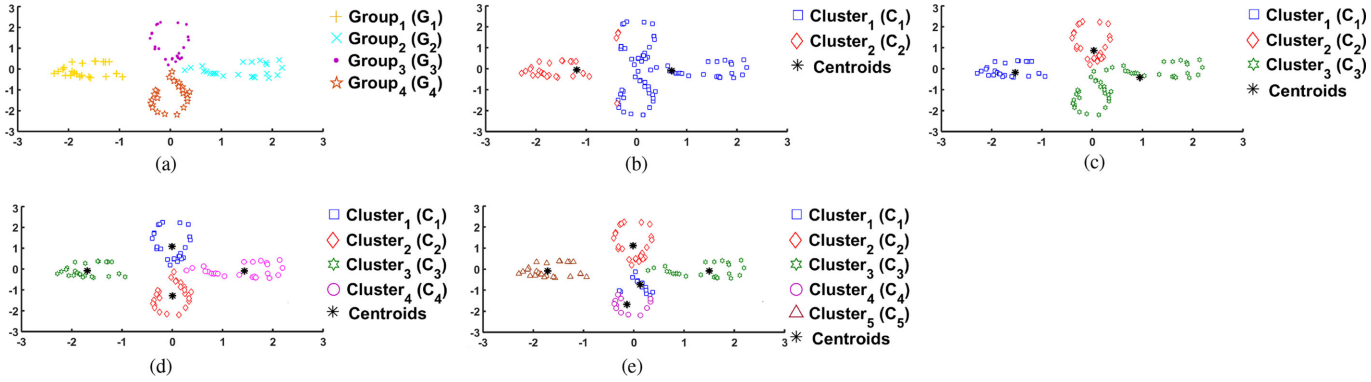

Fig. 1. Artificial petals data distribution [5] and corresponding clusters formed by FCM for different K values. (a) Ideal data. (b) $K = 2$. (c) $K = 3$. (d) $K = 4$. (e) $K = 5$.

TABLE I
EXISTING CVIS

S.No.	CVI	Formulation
1.	Dunn Index $(Dunn^+)$	$Dunn(K) = min_{1\le s\le K}\left(min_{s+1\le t\le K-1}\left(\frac{min_{x_t\in C_s, x_j\in C_t}(\|x_i - x_j\|)}{max_{1\le k\le K}\left[max_{x_t, x_j\in C_k}(\|x_i - x_j\|)\right]}\right)\right)$
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^+)	$CHI(K) = \frac{\left[\sum_{k=1}^{K}\|C_k\|\,\|v_k - \bar{v}\|^2\right]}{K-1} \Big/ \frac{\left[\sum_{k=1}^{K}\sum_{x_t\in C_k}\|x_i - v_k\|^2\right]}{N-K}$
4.	Davies and Bouldin Index (DBI^-)	$DBI(K) = \frac{1}{K}\sum_{k=1}^{K} max_{j\ne k}\left\{\frac{\left[\frac{1}{\lvert C_j\rvert}\sum_{x_i\in C_j}\|x_i - v_j\|^2\right] + \left[\frac{1}{\lvert C_k\rvert}\sum_{x_i\in C_k}\|x_i - v_k\|^2\right]}{\|v_j - v_k\|^2}\right\}$
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 $(PBMF^+)$	$PBMF(K) = \frac{1}{K}\times\frac{max_{j\ne k}\{\|v_j - v_k\|\}}{\sum_{k=1}^{K}\sum_{i=1}^{N}\mu_{ik}^m\|x_i - v_k\|}\times\sum_{i=1}^{N}\mu_{i1}\|x_i - v_1\|$
8.	PCAES Index $(PCAES^+)$	$PCAES(K) = \sum_{k=1}^{K}\sum_{i=1}^{N}\frac{\mu_{ik}^2}{\left[min_{1\le k\le K}\{\sum_{i=1}^{N}\mu_{ik}^2\}\right]} - exp\left(\frac{-min_{h\ne k}\{\|v_k - v_h\|^2\}}{\left[\frac{1}{K}\sum_{k=1}^{K}\|v_k - \bar{v}\|^2\right]}\right)$
9.	Wu-and-Li Index (WLI^-)	$WLI(K) = \frac{\left[\sum_{k=1}^{K}\left(\frac{\sum_{i=1}^{N}\mu_{ik}^2\|x_i - v_k\|^2}{\sum_{i=1}^{N}\mu_{ik}}\right)\right]}{2\times\left[\frac{1}{2}(min_{i\ne j}\{\|v_i - v_j\|^2\} + median_{i\ne j}\{\|v_i - v_j\|^2\})\right]}$
10.	V_R Index (V_R^-)	$V_R(K) = \sum_{k=1}^{c}\frac{(1/n_k)\sum_{ik}^{m}\|x_i - v_k\|^2 + (1/c)\|v_k - \bar{v}\|^2}{(1/(c-1))\sum_{j=1}^{c}\|v_j - v_k\|^2}$

compactness and intercluster separation. Wu and Yang [16] integrated the normalized PC with exponential separation to evaluate clustering and called it as the PC and 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], whereas S-index uses hierarchical-based clustering [19]. Furthermore, Sledge *et al.* [20] applied general CVIs on relational datasets. Moreover, clustering methods, such as 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 hyperellipsoid or hyperspherical 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, whereas G_1 is well separated from others. Furthermore, the clustering results for $K = 2$ and $K = 3$ are inappropriate, whereas $K = 5$ 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

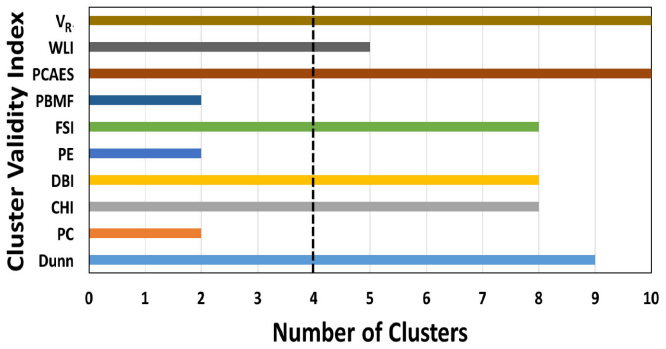

Fig. 2. Optimal cluster number suggested by different CVIs on the petals data distribution, as depicted in Fig. 1(a).

distribution. Fig. 2 reports the optimal cluster number suggested by the considered CVIs for $K = \{2, 3, \ldots, 9, 10\}$. In the figure, the black dotted line illustrates the correct cluster number for this distribution. It is visible from Fig. 2 that no 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 earlier.

In order to efficiently analyze hyperellipsoid or hyperspherical shape close clusters with distant centroids, this article 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 this article 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. Section V analyzes the experimental results. Finally, Section VI concludes this article.

II. Preliminaries

In this section, the FCM is explained first, followed by a brief description of the existing CVIs for comparative analysis. The notations used in the descriptions are listed as follows.

1) X: Dataset having N data points $X = \{x_1, x_2, \ldots, x_N\}$.
2) x_i: ith data point.
3) m: Level of cluster fuzziness.
4) K: Number of clusters.
5) C_k: kth cluster.
6) $\mid C_k \mid$: Total number of data points in kth cluster.
7) v_k: kth 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 ith data point with kth cluster.

A. Fuzzy C-Means

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 as

$$\sum_{i=1}^{N}\sum_{k=1}^{K} \mu_{ik}^{m} \parallel x_i - v_k \parallel^2, m \geq 1 \tag{1}$$

where $\mu_{ik} \in [0, 1]$ defines the degree of membership of ith data point with kth cluster. m is the fuzzification parameter that 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 the following:

$$U = [\mu_{ik}] \in \mathbb{R}^{N\times K},\ i = \{1, \ldots, N\},\ k = \{1, \ldots, K\}. \tag{2}$$

FCM optimizes (1) by updating μ_{ik} and v_k iteratively, which are defined as follows:

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

$$v_k = \frac{\sum_{i=1}^{N} \mu_{ik}^{m} x_i}{\sum_{i=1}^{N} \mu_{ik}^{m}}. \tag{4}$$

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

B. Existing CVI

A CVI purports to determine the correct clusters in X. The general procedure of a CVI is as follows.

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

For performance comparison, ten 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 intracluster 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 intercluster 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, whereas PC, PE, FSI, PCAES, PBMF, 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 $\bar{v}$, whereas 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 CVI

In this article, a new CVI, 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 the following:

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

where $\text{Co}(K)$ and $S(K)$ are fuzzy compactness and separation values for K clusters, calculated by (6) and (7), respectively

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

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

where

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

As observed from (6), compactness computes the fuzzy intracluster 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. Furthermore, (7) represents the separation as the intercluster variation in a K cluster system and considers the minimum distance among all the clusters based on data points, such as 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, as presented in Fig. 1(a). From this 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.

Furthermore, 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 in the following.

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

$$I \le \frac{1}{(D_1)^2}. \tag{9}$$

TABLE II
Result Analysis of SMI on the Petals Data Distribution, as 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^-.

It has already been proved by Dunn [10] that for hard clusters, the value of D_1 is always greater than one. Therefore, the aforementioned 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 article.

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

$$\text{Co}_{(\text{opt})} = (K-1)\left[\max_{\{1\le k\le K\}}\left(\frac{\sum_{i=1}^{N}\mu_{ik}^2||x_i - v_k||^2}{\sum_{i=1}^{N}\mu_{ik}}\right)\right]. \tag{10}$$

However, in case of hard K-partition, the corresponding compactness can be formulated as

$$\text{Co}_{(h)} = (K-1)\left[\max_{\{1\le k\le K\}}\left(\frac{\sum_{x_i\in C_k}||x_i - v_k||^2}{|C_k|}\right)\right]. \tag{11}$$

Ideally, it can be stated that

$$\text{Co}_{(\text{opt})} \le \text{Co}_{(h)}. \tag{12}$$

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

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

where $\text{dia}(C_k)$ is defined as

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

Thus, (12) can be rewritten as

$$\text{Co}_{(\text{opt})} \le (K-1)\left[\max_{\{1\le k\le K\}}\left(\frac{\sum_{x_i\in C_k}\text{dia}^2(C_k)}{|C_k|}\right)\right] \tag{15}$$

which can further be simplified as (16)

$$\text{Co}_{(\text{opt})} \le (K-1)\left[\max_{\{1\le k\le K\}}\{\text{dia}^2(C_k)\}\right]. \tag{16}$$

TABLE III
DETAILS OF CONSIDERED 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

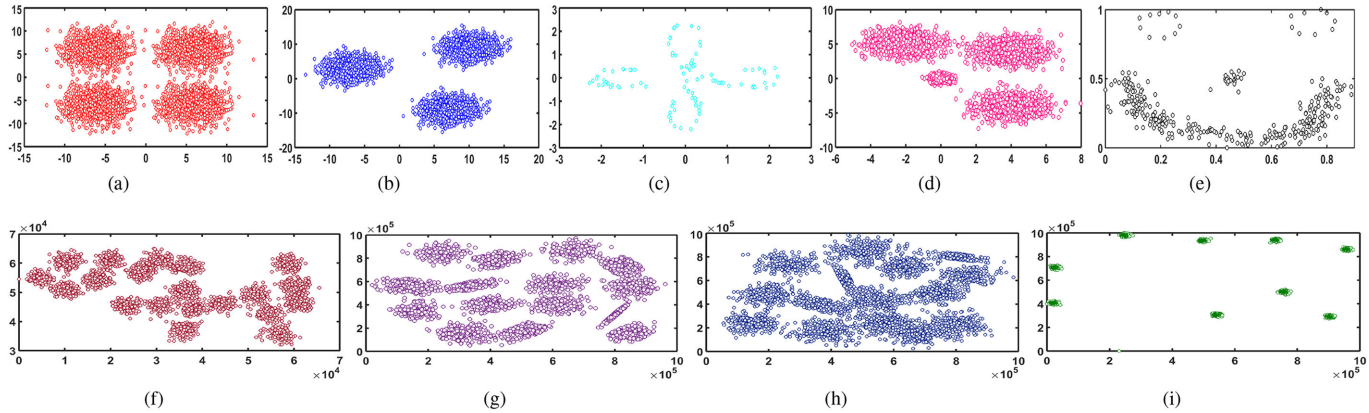

Fig. 3. Data distribution of ADs. (a) AD1. (b) AD2. (c) AD3. (d) AD4. (e) AD5. (f) AD6. (g) AD7. (h) AD8. (i) AD10.

Furthermore, as depicted in (7), the separation parameter (S) is represented by

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

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

$$\text{SMI} \le \frac{\text{Co}_{(\text{opt})}}{S} \tag{18}$$

$$\text{SMI} \le \frac{(K-1) * \max_{\{1\le k\le K\}} \text{dia}^2(C_k)}{\min_{\{1\le s\le K\}} \left(\min_{\{s+1\le t\le K-1\}} \left(\text{dist}^2(C_s, C_t) \right) \right)}. \tag{19}$$

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

$$\text{SMI} \le \frac{1}{(D_1)^2}. \tag{20}$$

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 (ADs) (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 ADs that 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, whereas AD3 to AD5 datasets are the standard synthetic datasets taken from the work in [26]. Furthermore, AD6 consists of a large number of circular clusters, whereas 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 ADs (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 colored 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 $K \in \{2, 3, \ldots, k, \ldots, 10\}$,

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

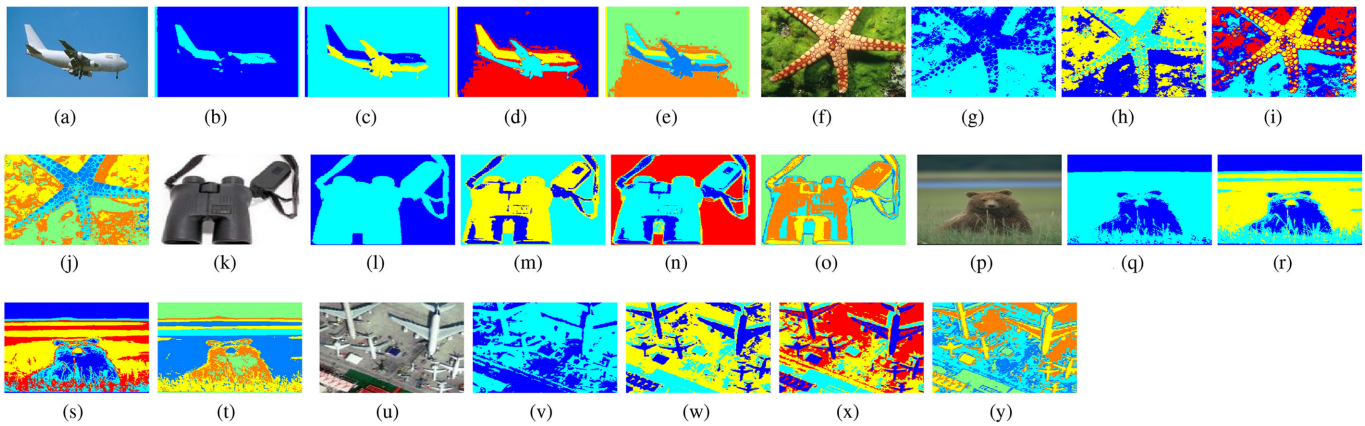

Fig. 4. Representative images (Img1–Img5) with clustering results on permissible K values (each color denotes a cluster). (a) Img1: Original. (b) Img1: $K = 2$. (c) Img1: $K = 3$. (d) Img1: $K = 4$. (e) Img1: $K = 5$. (f) Img2: Original. (g) Img2: $K = 2$. (h) Img2: $K = 3$. (i) Img2: $K = 4$. (j) Img2: $K = 5$. (k) Img3: Original. (l) Img3: $K = 2$. (m) Img3: $K = 3$. (n) Img3: $K = 4$. (o) Img3: $K = 5$. (p) Img4: Original. (q) Img4: $K = 2$. (r) Img4: $K = 3$. (s) Img4: $K = 4$. (t) Img4: $K = 5$. (u) Img5: Original. (v) Img5: $K = 2$. (w) Img5: $K = 3$. (x) Img5: $K = 4$. (y) Img5: $K = 5$.

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

$$K_{\text{best}} = \arg \max_{2 \le k \le 10} O(k) \tag{21}$$

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

After all the rounds, the sensitivity of each CVI for a dataset is also analyzed. The formulation of sensitivity is defined as

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

where $\#C$ is the true cluster numbers for a dataset and count$(k = \#C)$ refers to the number of times k for a CVI matches $\#C$ in 50 rounds. Furthermore, 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 $a \le b$ (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

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

where

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

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

V. EXPERIMENTAL RESULTS

Table V illustrates the results of the proposed and considered CVIs for ADs. In Table V(a), K_{best} value of SMI equals $\#C$ for all ADs 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 this figure, it is observable that FCM is biased toward 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 AD (AD3), only the proposed SMI gives accurate results with the sensitivity of 0.72 and rest CVIs do not suggest accurate cluster numbers even for a single time. It illustrates the efficiency of both compactness and separation components of SMI. Furthermore, only Dunn and SMI determine the correct clusters for AD6. On AD7, PE, PBMF, PCAES,

TABLE V
CLUSTERING RESULTS ON ADS 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

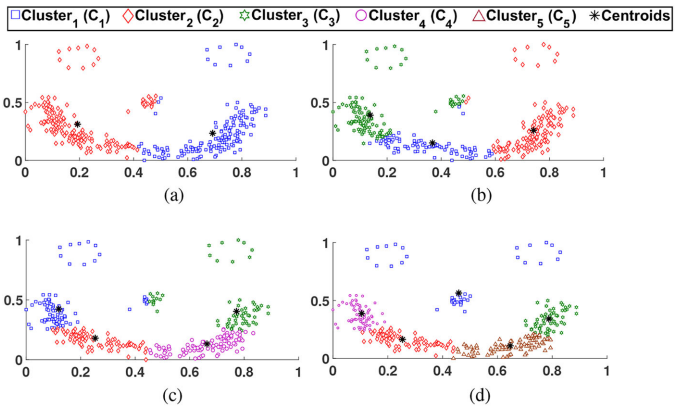

Fig. 5. Clustering results of FCM on AD5 for different K values. (a) $K = 2$. (b) $K = 3$. (c) $K = 4$. (d) $K = 5$.

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 this 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.

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

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, for all UCIs except UCI6 for which it returns 0.68. For the UCI1 dataset, only Dunn, FSI, and SMI give competitive results. Furthermore, PCAES does not return the correct cluster number for any UCI dataset. Moreover, FSI fails on every UCI dataset except UCI1, whereas 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.

Furthermore, the clustering results of the FCM method on the image dataset for $K \in \{2, 3, \ldots, 5\}$ 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 this 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. Furthermore, 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., aircraft, are distinctly segmented. Moreover, Table VII(b) illustrates the average sensitivity of the considered CVIs. SMI attains the average sensitivity of 1, 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 ADs, the proposed CVI attains the maximum effectiveness, i.e., 0.95. Likewise, for both UCI and image datasets, only SMI has 1 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

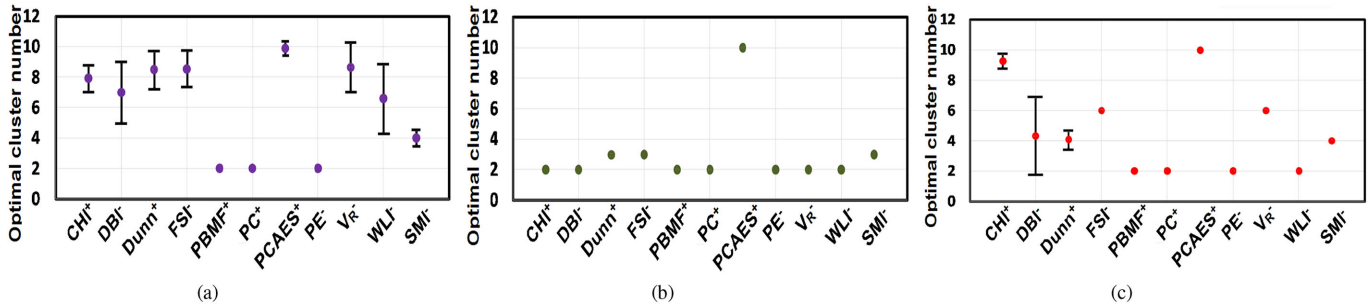

Fig. 6. Error-bar plots with standard deviation of considered CVIs on representative dataset of (a) AD1, (b) UCI dataset (UCI1), and (c) image dataset (Img5).

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 article introduced a new CVI, SMI, for hyperellipsoid or hyperspherical 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 was favorable for handling the closely allocated clusters with distant centroids. The experimental evaluation manifested that SMI performed better and more stable than the ten state-of-the-art considered CVIs on artificial, UCI, and image datasets. Some worthwhile aspects of evaluations are as follows.

1) SMI was efficient in reporting optimal clusters for all ADs except AD5. However, none of the CVI reported correct cluster numbers for AD5 due to the presence of a very high dense and nonspherical cluster as compared to the other three sparse clusters. It also affected the clustering results of the FCM method. Correspondingly, it was reasonable to consider the splitting of AD5 into two clusters as done by proposed SMI.
2) The considered UCI datasets were the standard clustering datasets with the varied number of features on which SMI outperformed all the considered CVIs. Furthermore, it was also interesting to note the stability of the results, produced by SMI, which signified that the cluster number suggested by SMI was either optimum or near to optimum in each round.
3) The performance of SMI was further evaluated on pixel intensities of RGB colored images, having a range of optimal clusters as augmented by domain experts. Although the shape of a multidimensional dataset is infeasible to determine and state whether the distribution is hyperellipsoid or hyperspherical 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.

1) The performance of SMI on noisy datasets and clustering applications is yet to be witnessed. Moreover, clustering methods other than centroid-based clustering, such as 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.
2) 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.
3) 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.
4) 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, such as pixel position or texture.

5) 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.

REFERENCES

[1] J.-Y. Jiang, R.-J. Liou, and S.-J. Lee, "A fuzzy self-constructing feature clustering algorithm for text classification," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 3, pp. 335–349, Mar. 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 Trans. Fuzzy Syst.*, vol. 23, no. 3, pp. 701–718, Jun. 2015.

[3] M. K. Pakhira, S. Bandyopadhyay, and U. Maulik, "Validity index for crisp and fuzzy clusters," *Pattern Recognit.*, vol. 37, no. 3, pp. 487–501, Mar. 2004.

[4] E. Hancer and D. Karaboga, "A comprehensive survey of traditional, merge-split and evolutionary approaches proposed for determination of cluster number," *Swarm. Evol. Comput.*, 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 Trans. Fuzzy Syst.*, vol. 19, no. 1, pp. 152–163, Feb. 2011.

[6] J. C. Bezdek, "Cluster validity with fuzzy sets," *J. Cybern.*, vol. 3, pp. 58–73, 1973.

[7] N. R. Pal and J. C. Bezdek, "On cluster validity for the fuzzy c-means model," *IEEE Trans. Fuzzy Syst.*, vol. 3, no. 3, pp. 370–379, Aug. 1995.

[8] M.-C. Chiang, C.-W. Tsai, and C.-S. Yang, "A time-efficient pattern reduction algorithm for k-means clustering," *Inf. Sci.*, 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 Res. Inst., Menlo Park, CA, USA, Tech. Rep. AD699616, 1965.

[10] J. C. Dunn, "A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters," *J. Cybern.*, vol. 3, pp. 32–57, 1973.

[11] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Commun. Statist.-Theory Methods*, vol. 3, pp. 1–27, 1974.

[12] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979.

[13] J. C. Bezdek, "Objective function clustering," in *Pattern Recognition with Fuzzy Objective Function Algorithms*. Berlin, Germany: 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. Fuzzy Syst. Symp.*, 1989, pp. 247–250.

[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 Syst.*, vol. 155, pp. 191–214, 2005.

[16] K.-L. Wu and M.-S. Yang, "A cluster validity index for fuzzy clustering," *Pattern Recognit. 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," *Comput. Intell. Neurosci.*, 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 Trans. Syst., Man, Cybern., Part B (Cybern.)*, vol. 41, no. 4, pp. 1039–1053, Aug. 2011.

[19] Z. Liang, P. Zhang, and J. Zhao, "Optimization of the number of clusters in fuzzy clustering," in *Proc. Int. Conf. Comput. Des. Appl.*, 2010, pp. 580–584.

[20] I. J. Sledge, J. C. Bezdek, T. C. Havens, and J. M. Keller, "Relational generalizations of cluster validity indices," *IEEE Trans. Fuzzy Syst.*, vol. 18, no. 4, pp. 771–786, Aug. 2010.

[21] A. Jose-Garcia and W. Gómez-Flores, "Automatic clustering using nature-inspired metaheuristics: A survey," *Appl. Soft Comput.*, 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 Recognit. Lett.*, 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. Int. Conf. Fuzzy Syst.*, 2014, pp. 1526–1533.

[24] X. L. Xie and G. Beni, "A validity measure for fuzzy clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 8, pp. 841–847, Aug. 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." Accessed: Jan. 23, 2018, 2018. [Online]. Available: https://in.mathworks.com/matlabcentral/fileexchange/41459-6-functions-for-generating-artificial-datasets

[27] P. Fränti and S. Sieranoja, "K-means properties on six clustering benchmark datasets," *Appl. Intell.*, vol. 48, pp. 4743–4759, 2018. [Online]. Available: http://cs.uef.fi/sipu/datasets/

[28] "UCI machine learning repository: Data sets." Accessed: Jan. 23, 2018, 2018. [Online]. Available: http://archive.ics.uci.edu/ml/datasets.html

[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. Int. Conf. Comput. Vis.*, 2001, pp. 416–423.

[30] M. R. Anderberg, *Cluster Analysis for Applications: Probability and Mathematical Statistics: A Series of Monographs and Textbooks*, 1st ed. Orlando, FL, USA: Academic, 2014.

Himanshu Mittal was born in 1988. He received the M.Tech. degree in software engineering from Delhi Technological University, New Delhi, India, in 2012, and the Ph.D. degree from the Department of Computer Science and Engineering, Jaypee Institute of Information Technology, Noida, India, in 2020.

He has more than seven years of teaching and research experience. He is currently a part of SERB-DST (New Delhi) funded project on Histopathological Image Analysis. He has authored or coauthored more than 20 journal and conference papers in the area of image processing, pattern recognition, data mining, and soft computing. His current research interests include image processing, pattern recognition, and soft computing.

Mukesh Saraswat was born in 1978. He received the Ph.D. degree in computer science & engineering from Atal Bihari Vajpayee-Indian Institute of Information Technology and Management Gwalior, Gwalior, India, in 2013.

He is currently an Associate Professor with the Jaypee Institute of Information Technology, Noida, India. He has more than 18 years of teaching and research experience. He has guided two Ph.D. students and currently guiding four Ph.D. students. He was part of the successfully completed DRDE funded project on image analysis and the currently running two projects funded by SERB-DST (New Delhi) on Histopathological Image Analysis and Collaborative Research Scheme, under TEQIP III (RTU-ATU) on Smile. He has authored or coauthored more than 40 journal and conference papers in the area of image processing, pattern recognition, data mining, and soft computing. His current research interests include image processing, pattern recognition, mining, and soft computing.

Dr. Saraswat has been an active member of many organizing committees of various conferences and workshops. He was also a Guest Editor for the *Journal of Swarm Intelligence*. He is an active member of ACM and CSI professional bodies.