S-Divergence-Based Internal Clustering Validation Index

Krishna Kumar Sharma[1], Ayan Seal[2,6]*, Anis Yazidi[3,4,5], Ondrej Krejcar[6,7]

[1] Department of Computer Science and Informatics, University of Kota, Kota, Rajasthan-324005 (India)
[2] Department of Computer Science and Engineering, PDPM Indian Institute of Information Technology Design & Manufacturing Jabalpur, Jabalpur, Madhya Pradesh-482005 (India)
[3] Department of Computer Science, OsloMet–Oslo Metropolitan University, Oslo, 460167, (Norway)
[4] Department of Computer Science, Norwegian University of Science and Technology, Trondheim, 460167, (Norway)
[5] Department of Plastic and Reconstructive Surgery, Oslo University Hospital, Oslo, 460167, (Norway)
[6] Center for Basic and Applied Science, Faculty of informatics and management, University of Hradec Kralove, Rokitanskeho 62, 50003 Hradec Kralove, (Czech Republic)
[7] Malaysia-Japan International Institute of Technology (MJIIT), Universiti Teknologi Malaysia, Jalan Sultan Yahya Petra, 54100 Kuala Lumpur, (Malaysia)

Received 22 July 2022 | Accepted 14 January 2023 | Published 24 October 2023

Abstract

A clustering validation index (CVI) is employed to evaluate an algorithm's clustering results. Generally, CVI statistics can be split into three classes, namely internal, external, and relative cluster validations. Most of the existing internal CVIs were designed based on compactness (CM) and separation (SM). The distance between cluster centers is calculated by SM, whereas the CM measures the variance of the cluster. However, the SM between groups is not always captured accurately in highly overlapping classes. In this article, we devise a novel internal CVI that can be regarded as a complementary measure to the landscape of available internal CVIs. Initially, a database's clusters are modeled as a non-parametric density function estimated using kernel density estimation. Then the S-divergence (SD) and S-distance are introduced for measuring the SM and the CM, respectively. The SD is defined based on the concept of Hermitian positive definite matrices applied to density functions. The proposed internal CVI (PM) is the ratio of CM to SM. The PM outperforms the legacy measures presented in the literature on both superficial and realistic databases in various scenarios, according to empirical results from four popular clustering algorithms, including fuzzy k-means, spectral clustering, density peak clustering, and density-based spatial clustering applied to noisy data.

Keywords

Cluster Validity Index, Generalized Mean, *K*-nearest Neighbors, S-distance, S-divergence, Spectral Clustering, Symmetry Favored.

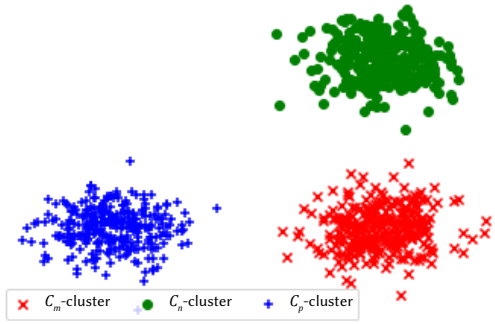
DOI: 10.9781/ijimai.2023.10.001

I. Introduction

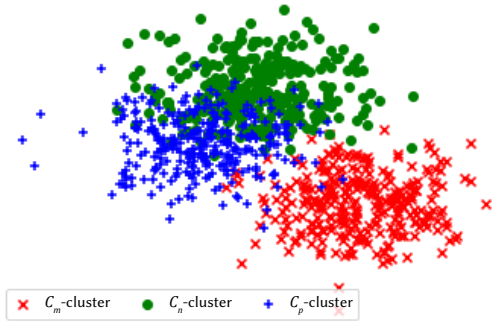
Clustering is an unsupervised methodology for analyzing a set of data objects by dividing them into subsets such that each group contains similar objects while dissimilar ones end up in different groups [1]–[5]. Thus, the objective of clustering is to mine the data to explore multi-dimensional obscure patterns and hidden structures in the data. Nowadays, clustering has received a great deal of attention among the community of researchers in the area of pattern recognition by the virtue of remarkable academic and commercial applications spanning over a wide range which includes identifying fake news [6], spam filtering [7], market segmentation [8], [9], classifying network traffic [10], detecting fraudulent or criminal activity [11], [12], cybersecurity [13], document analysis [14], drug discovery [15], information retrieval [16], and many more [17]–[22].

* Corresponding author.
E-mail address: ayan@iiitdmj.ac.in

A fundamental question in clustering is how to assess the "goodness" of the resulting clusters. The answer to this question is not obvious as it is difficult to devise criteria that determine the optimal partitioning of the data objects into clusters. Obtaining insights about the goodness of clusters using some visualization tools is not a feasible solution when the number of dimensions increases, as human eyes are not accustomed to higher-dimensional spaces. The process of assessing the performance of the clustering algorithm is referred to as cluster validation. According to the clustering validation procedure, the outcome of the clustering phase is validated quantitatively by a Clustering Validation Index (CVI). A CVI can be considered a function that, for a given clustering scheme and database, produces some value that represents the quality of the clustering scheme [23], [24]. In other words, a CVI provides some insight into the quality of grouping. Internal, external, and relative are the three main categories of CVIs. Internal CVIs rely only on the internal information of a given database. Unlike internal CVIs, external CVIs assess the "goodness" of a clustering structure based on provided class labels as external inputs [25]–[28]. On the other hand, relative CVIs evaluate the clustering



(a) An instance of well separated clusters



(b) An instance of overlapped clusters

Fig. 1. Distribution of three clusters.

results by changing the number of clusters. We mainly concentrate on internal CVIs in this work. The most intuitive notions for defining "good clustering" are cohesion/compactness (CM) and separation (SM). In simple words, when data objects in a cluster are in the vicinity of each other, the cluster is called a compact cluster. On the other hand, when neighboring clusters are possibly quite far from each other, then these clusters are easily identifiable and well separated. In other words, SM measures the distance between the centers of two clusters, whereas CM measures the variance within a cluster. Generally, geometric distance is used to compute SM. However, geometric distance can not always represent the SM efficiently, especially when two clusters are highly overlapping. Let us consider an example where three clusters, namely C_m, C_n, and C_p, are well separated (see Fig. 1(a)). As clusters are well separated, geometric distance can efficiently capture the dissimilarity between clusters. We may assume another scenario (see Fig. 1(b)), where clusters C_m, C_n, and C_p, are overlapping. In this case, the geometric distance between the centers of C_m and C_n is the same as the geometric distance between the centers of C_m and C_p. Thus, the dissimilarity between clusters can not be captured accurately using geometric distance. In [29], Cui et al. assumed that the data of a cluster were obtained from multivariate Gaussian distributions, and Jeffrey divergence (JD) was considered, as a distance measure for computing SM between clusters. The JD is not a valid distance measure because it does not abide by the metric property of triangle inequality [30]. In addition, the JD is not appropriate, while clusters are almost identical. Alternatively stated, a small change in clusters cannot be captured by JD. It encourages us to delve further in this direction by proposing an internal CVI based on the notion of S-divergence (SD), which can catch tiny variations in clusters since the cone is formed by the Hermitian positive definite matrices (HPDM). In addition, it fulfills all the properties of the distance metric [31]. Four well-known clustering techniques are employed to evaluate the performance of the proposed CVI on ten real-world and artificial databases. However, each cluster is modeled as a random variable using a non-parametric probability density function named kernel density estimation (KDE) before the use of the proposed internal CVI. Among ten databases, some are well separated, a few are slightly overlapping, and the rest are highly overlapping. Moreover, noise is added in some databases to validate the efficacy of the proposed CVI. A comparative analysis is also performed to show the competitiveness of the proposed internal CVI in comparison with other CVIs.

The remaining article is structured as follows. After the introduction, in Section II, we examine several well-known internal CVIs. The proposed internal CVI is discussed further in Section III. Section IV provides the results of the experiments. At last, Section V concludes the work.

II. Related Works

A summary of some of the most popular internal CVIs is presented in this section. The CM reflects the average closeness or similarity of data points in all clusters. A value approaching 0 indicates good clustering [32]. The SM portrays the degree of separation between clusters [32]. A higher value of SM signifies better clustering. It is worth mentioning that other indexes, for example, root mean square standard deviation index (RMSSTDI) [33], root squared index (RSI) [33], and modified Hubert validity index (MHI) [34] perform on a different principle. Indeed, the RMSSTDI quantifies the homogeneity of the resultant clusters by calculating the square root of the aggregated variance of all the data objects. RSI determines the magnitude of difference between clusters using the ratio of the addition of the squares **between-clusters** to the total summation of the squares in the database. RMSSTDI, RSI, and MHI evaluate the difference **between-clusters** by calculating the disagreements of groups of data objects in two parts. Furthermore, these indexes do not consider both CM and SM to validate the formed clusters. The Calinski-Harabasz index (CHI) computes the ratio of the sum of the average of **between-clusters** and of **intra-cluster dispersion** for all clusters [35]. A greater value of CHI demonstrates better partitions. CHI is usually fast to compute. Moreover, it is suitable for convex and well-separated clusters. On the other hand, it produces a low value for non-convex clusters. The Dunn validity index (DVI) calculates the SM of clusters over the CM of clusters [36]. Thus, a larger value of DVI suggests well-separated and compact clusters. However, the complexity of the DVI increases with the increase in the number of clusters, k. The Davies-Bouldin index (DBI) computes cluster overlapping using the ratio of the sum of **intra-cluster** spread to **between-cluster** distance [37]. A value adjacent to 0 illustrates better partitions. It computes the inherent attributes and quantities of a database. Moreover, it is limited to Euclidean space. The JD-based validity index (JI) is a ratio of CM to JD-based SM, [38]. JD determines the similarity between two probability distributions and is suitable for slightly-overlapping clusters. Thus, a value close to 0 is a sign of better partitions. However, JD falls short when the clusters are highly overlapping. The silhouette index (SI) measures how alike a data object is to its own cluster/cohesion/CM against other clusters/SM [39]. A value near 1 signifies that the data object is well-suited to its cluster and does not match enough to neighboring clusters. A clustering configuration is appropriate when most data objects have a high value. SI is higher for well-separated and dense clusters. However, it is not suitable for non-convex clusters. Moreover, the computational complexity, $O(n^2 d\ log(n))$, is high. I validity index (IVI) computes the CM and the SM using the maximum distance among data objects and centers of clusters [40]. Furthermore, the optimal number of clusters is calculated by maximizing the value of IVI. The Xie-Beni index (XBI) is defined using CM as the mean square distance among data objects and their cluster centers and the SM as the minimum square distance between the centers of clusters [41]. Optimal clusters exhibit a minimum value of XBI. The value of XBI reduces monotonically as the value of k increases. Furthermore, Bouguessa et al. [42] and Arbelaitz

TABLE I. A Review of Some of the Popular Internal CVIs

S. No.	Internal CVI	Notation	Expression	Range	Optimal value	Complexity
1	Root mean square standard deviation index	RMSSTDI	$\left\{\frac{\sum_{i=1}^{k}\sum_{c_j\in C_i}\|\mathbf{c}_j-\mathbf{v}_i\|^2}{d\sum_{i=1}^{k}(\|C_i\|-1)}\right\}^{1/2}$	[0, +∞]	elbow	$O(nd)$
2	Root squared index	RSI	$\frac{\sum_{\mathbf{c}\in DB}\|\mathbf{c}-\mathbf{v}\|^2-\sum_{i=1}^{k}\sum_{c_j\in C_i}\|\mathbf{c}_j-\mathbf{v}_i\|^2}{\sum_{\mathbf{c}\in DB}\|\mathbf{c}-\mathbf{v}\|^2}$	[0, 1]	elbow	$O(nd)$
3	Modified Hubert validity index	MHI	$\frac{2}{n(n-1)}\sum_{\mathbf{c}_j\in C_i \text{ and } \mathbf{v}_i\in C_i}\ \sum_{\mathbf{c}_q\in C_r \text{ and } \mathbf{v}_r\in C_r}\text{dist}(\mathbf{c}_j,\mathbf{c}_q)\text{dist}(\mathbf{v}_i,\mathbf{v}_r)$	[0, +∞]	elbow	$O(n^2d)$
4	Compactness measure	CM	$\frac{1}{k}\sum_{i=1}^{k}\frac{1}{\|C_i\|}\sum_{c_j\in C_i}dist(\mathbf{c}_j,\mathbf{v}_i)$	[0, +∞]	Min	$O(nd)$
5	Separation measure	SM	$\frac{2}{k^2-k}\sum_{i=1}^{k}\sum_{p=i+1}^{k}\text{dist}(\mathbf{v}_i,\mathbf{v}_p)$	[0, +∞]	Max	$O(k^2d)$
6	Calinski-Harabasz index	CHI	$\sum_{i=1}^{k}\frac{\|C_i\|\times\frac{dist(\mathbf{v}_i,\mathbf{v})}{(k-1)}}{\sum_{c_j\in C_i}\frac{dist(\mathbf{c}_j,\mathbf{v}_i)}{(n-k)}}$	[0, +∞]	Max	$O(nd)$
7	Dunn validity index	DVI	$\frac{\min_{1\leqslant i\neq r\leqslant k}\left(\min_{\forall c_{c_0}\in C_0,\forall c_q\in C_r}\{\text{dist}(\mathbf{c}_j,\mathbf{c}_q)\}\right)}{\forall c_{c_j}\in C_{i,j},\forall c_q\in C_r}$	[0, +∞]	Max	$O(n^2d\log(n))$
8	Davies-Bouldin index	DBI	$\frac{1}{k}\sum_{i=1}^{k}\max_{r\neq i}\left(\frac{\frac{1}{\|C_i\|}\sum_{c_j\in C_i}\text{dist}(\mathbf{c}_j,\mathbf{v}_i)+\frac{1}{\|C_l\|}\sum_{c_j\in C_p}\text{dist}(\mathbf{c}_j,\mathbf{v}_p)}{\text{dist}(\mathbf{v}_i,\mathbf{v}_p)}\right)$	[0, +∞]	Min	$O(n^2d\log(n))$
9	Jeffrey-divergence based validity index	JI	$\frac{\frac{1}{k}\sum_{i=1}^{k}\frac{1}{\|C_i\|}\sum_{c_j\in C_i}dist(\mathbf{c}_j,\mathbf{v}_i)}{\frac{2}{k^2-k}\sum_{i=1}^{k}\sum_{p=i+1}^{k}JD(\mathbf{v}_i,\mathbf{v}_p)}$	[0, +∞]	Min	$O(nd)$
10	Silhouette index	SI	$\frac{1}{n}\sum_{i=1}^{k}\sum_{c_j\in C_i}\frac{\text{sep}(\mathbf{c}_j,\mathbf{c}_q)-\text{coh}(\mathbf{c}_j,\mathbf{c}_i)}{\max\{\text{sep}(\mathbf{c}_j,\mathbf{c}_q),\text{coh}(\mathbf{c}_j,\mathbf{c}_i)\}}$, where $\text{coh}(\mathbf{c}_j,\mathbf{c}_i)=\frac{1}{\|C_i\|}\sum_{c_i\in C_i}\text{dist}(\mathbf{c}_j,\mathbf{c}_i)$ and $\text{sep}(\mathbf{c}_j,\mathbf{c}_q)=\min_{C_r\neq i \text{ and } 1\leqslant r\leqslant k}\frac{1}{\|C_r\|}\sum_{c_q\in C_r}\text{dist}(\mathbf{c}_j,\mathbf{c}_q)$	[-1, 1]	Max	$O(n^2d\log(n))$
11	I validity index	IVI	$\left(\frac{1}{k}\times\frac{\sum_{\mathbf{c}\in DB}\text{dist}(\mathbf{c},\mathbf{v})}{\sum_{i=1}^{k}\sum_{c_j\in C_i}\text{dist}(\mathbf{c}_j,\mathbf{v}_i)}\times\max_{1\leqslant j,q\leqslant n}\{\text{dist}(\mathbf{c}_j,\mathbf{c}_q)\}\right)^d$	[0, +∞]	Max	$O(n^2d\log(n))$
12	Xie-Beni index	XBI	$\sum_{i=1}^{k}\frac{\sum_{c_j\in C_i}\text{dist}^2(\mathbf{c}_j,\mathbf{v}_i)}{n\times\min_{\mathbf{c}_j,\mathbf{c}_q\neq c_j,}\text{dist}^2(\mathbf{c}_j,\mathbf{c}_q)}$	[0, +∞]	Min	$O(n^2d\log(n))$

DB: Dataset, n: number of data objects in DB, v : center of DB, d: number of attributes, c: data objects of DB, k: number of clusters, C_i: i^{th} cluster, $\mathbf{c}_j$: j^{th} member of i^{th} cluster, $\mathbf{v}_i$: center of i^{th} cluster, $var(C_i)$: variance vector of C_i, dist ( ): distance function.

et al. [43] also worked to introduce indices based on Dunn variations and cohesion, which act well with noisy and overlapped clusters. Table I reports the definition, range, optimum value, and complexity of each of the above-discussed internal CVIs.

III. Proposed CVI

In this section, we examine and present some of the imperative properties of SD and propose a new internal CVI measure.

A. S-Divergence and Its Properties

Definition 1. SD presents a metric on the set of matrices A_τ of size $\tau\times\tau$ [31]. The set A_τ is a convex cone, on which SD is defined using Eq. 1.

$$D_S^2(A_\tau^i,A_\tau^j)=\log(\det(\frac{A_\tau^i+A_\tau^j}{2}))-\frac{1}{2}\log(\det(A_\tau^iA_\tau^j)) \tag{1}$$

where det(.) denotes the determinant operation. D_S is a metric on the positive definite matrices (PDM) A_τ. Let ϕ_τ be a one-to-one function from $\Re_+^\tau\to A_\tau$. Now examine a vector $\mathbf{t}=\{t_1,t_2,\ldots,t_\tau\}\in\Re_+^\tau$ to generate PDM from a vector $\mathbf{t}$. SD is a divergence function on the cone of HPDM. A convex cone structure on the set of HPDM enables "geometric optimization", which enables us to resolve certain problems that may be non-convex in Euclidean space but convex in manifold space, or, offers efficient optimization. Thus, the divergence function on the cone of hpd matrices has empirical and computational advantages in many applications [44].

At this juncture, we shall demonstrate that the SD meets all the necessary characteristics for becoming a distance metric, which are given below:

Proposition 1. *Non-negativity*: $D_S^2(\phi_\tau(\mathbf{t}),\phi_\tau(\mathbf{u}))\geq 0$

Proof. The modified version of Eq. 1 is given below:

$$D_S^2(\phi_\tau(\mathbf{t}),\phi_\tau(\mathbf{u}))=\log(\det(\frac{\phi_\tau(\mathbf{t})+\phi_\tau(\mathbf{u})}{2}))+\log(\frac{1}{\sqrt{\det(\phi_\tau(\mathbf{t})\phi_\tau(\mathbf{u}))}}) \tag{2}$$

$$\Rightarrow D_S^2(\phi_\tau(\mathbf{t}),\phi_\tau(\mathbf{u}))=\log(\frac{\det(\frac{\phi_\tau(\mathbf{t})+\phi_\tau(\mathbf{u})}{2})}{\sqrt{\det(\phi_\tau(\mathbf{t})\phi_\tau(\mathbf{u}))}}) \tag{3}$$

where $\frac{\det(\frac{\phi_\tau(\mathbf{t})+\phi_\tau(\mathbf{u})}{2})}{\sqrt{\det(\phi_\tau(\mathbf{t})\phi_\tau(\mathbf{u}))}}\geq 0$ because determinant of the PDM is always positive and numerator will be greater thanequal to denominator. $\therefore D_S^2(\phi_\tau(\mathbf{t}),\phi_\tau(\mathbf{u}))\geq 0$

□

Proposition 2. *Equality*: $D_S^2(\phi_\tau(\mathbf{t}),\phi_\tau(\mathbf{u}))=0$ *iff* $\mathbf{t}=\mathbf{u}$

Proof. From proposition 1, we can write

$$D_S^2(\phi_\tau(\mathbf{t}),\phi_\tau(\mathbf{u}))=\log(\frac{\det(\frac{\phi_\tau(\mathbf{t})+\phi_\tau(\mathbf{u})}{2})}{\sqrt{\det(\phi_\tau(\mathbf{t})\phi_\tau(\mathbf{u}))}})$$

Now, if **t** and **u** are equal then **u** can be replaced by **t** in the above expression and the modified expression is

$$D_S^2(\phi_\tau(\mathbf{t}),\phi_\tau(\mathbf{t})) = \log\left(\frac{\det\left(\frac{\phi_\tau(\mathbf{t})+\phi_\tau(\mathbf{t})}{2}\right)}{\sqrt{\det(\phi_\tau(\mathbf{t})\phi_\tau(\mathbf{t}))}}\right) \Rightarrow \log(1) = 0$$

$$\therefore D_S^2(\phi_\tau(\mathbf{t}),\phi_\tau(\mathbf{u})) = 0 \text{ iff } \mathbf{t} = \mathbf{u}.$$

Please note that we used the property that the determinant of the power of a matrix is equal to the determinant raised to that power, meaning in our case:

$$\det(\phi_\tau(\mathbf{t})\phi_\tau(\mathbf{t})) = \det(\phi_\tau(\mathbf{t}))^2$$

□

Proposition 3. *Symmetry*:
$D_S^2(\phi_\tau(\mathbf{t}),\phi_\tau(\mathbf{u})) = D_S^2(\phi_\tau(\mathbf{u}),\phi_\tau(\mathbf{t}))$

Proof. The SD amid **t** and **u** is denoted as follows:

$D_S^2(\phi_\tau(\mathbf{t}),\phi_\tau(\mathbf{u})) = \log\left(\frac{\det\left(\frac{\phi_\tau(\mathbf{t})+\phi_\tau(\mathbf{u})}{2}\right)}{\sqrt{\det(\phi_\tau(\mathbf{t})\phi_\tau(\mathbf{u}))}}\right)$ [as already noted in proposition 1] $= D_S^2(\phi_\tau(\mathbf{u}),\phi_\tau(\mathbf{t}))$

$\therefore D_S^2(\phi_\tau(\mathbf{t}),\phi_\tau(\mathbf{u})) = D_S^2(\phi_\tau(\mathbf{u}),\phi_\tau(\mathbf{t}))$

It implies SD also abides the symmetric metric property.

□

Proposition 4. *Triangle Inequality: Suppose* ***t****,* ***u****, and* ***z*** *be three vectors. Then this proposition states, the sum of the lengths of any two sides viz.,* $D_S(\phi_\tau(\boldsymbol{t}), \phi_\tau(\boldsymbol{u}))$ *and* $D_S(\phi_\tau(\boldsymbol{u}), \phi_\tau(\boldsymbol{z}))$ *of a triangle is greater than or equal to the length of the third side* $D_S(\phi_\tau(\boldsymbol{t}), \phi_\tau(\boldsymbol{z}))$*. Arithmetically,* $D_S(\phi_\tau(\boldsymbol{t}), \phi_\tau(\boldsymbol{z})) \leq D_S(\phi_\tau(\boldsymbol{t}), \phi_\tau(\boldsymbol{u})) + D_S(\phi_\tau(\boldsymbol{u}), \phi_\tau(\boldsymbol{z}))$.

Proof. Let **t**, **u**, and **z** be three vectors. Then $\phi_\tau(\boldsymbol{t})$, $\phi_\tau(\boldsymbol{u})$, $\phi_\tau(\boldsymbol{z}) > 0$ and diagonal matrices.

Thus $D_S^2(\phi_\tau(\mathbf{t}),\phi_\tau(\mathbf{u})) = \sum_i D_S^2(t_i, u_i)$,

$D_S^2(\phi_\tau(\mathbf{t}),\phi_\tau(\mathbf{z})) = \sum_i D_S^2(t_i, z_i)$, and

$$D_S^2(\phi_\tau(\mathbf{u}),\phi_\tau(\mathbf{z})) = \sum_i D_S^2(u_i, z_i)$$

$$\therefore D_S^2(\phi_\tau(\mathbf{t}),\phi_\tau(\mathbf{z})) \leq D_S^2(\phi_\tau(\mathbf{t}),\phi_\tau(\mathbf{u})) + D_S^2(\phi_\tau(\mathbf{u}),\phi_\tau(\mathbf{z}))$$

□

Hence, it is showed that the SD is a metric.

B. Cluster Density Estimation

In this study, each cluster is modeled using a random variable characterized by a probability distribution. In practice, the underlying probability distribution of a random variable is not known in advance. Alternatively, the probability distribution of a random variable is estimated from the data objects or samples of a cluster. Therefore, each random variable is associated with a set of samples. We assume that samples are finite, independent, and identically distributed. Here, we adopt the well-known non-parametric probability estimation technique KDE to estimate the underlying distribution of the observations.

Let M be a random variable characterizing cluster C_m, where each sample, $\mathbf{x}$, is of $d-$dimensions. Then, the kernel function is obtained by multiplying the d number of Gaussian functions with bandwidth, h_l^M, where $1 \leq l \leq d$ and $d \geq 2$. Equation 4 is applied to estimate M [1], [2].

$$M(\mathbf{x}) = \frac{1}{|C_m|(2\pi)^{\frac{d}{2}}\prod_{l=1}^{d} h_l^M} \sum_{\mathbf{c}_j \in C_m} \prod_{l=1}^{d} e^{-\frac{(x^l - c_{j,l})^2}{2h_l^{M^2}}} \tag{4}$$

where $x \in \mathcal{D}$, every cluster is defined in the same domain $\mathcal{D}$ and we also assume that the $\mathcal{D}$ is a bounded range of values and c_j is a j^{th} member of i^{th} cluster or $c_j \in C_i$. h_l^M is the bandwidth of the l^{th} feature and it controls the smoothing of the Gaussian kernel function. The Sliverman approximation rule (Eq. 5) is considered to estimate h_l^M.

$$h_l^M = 1.06 \times \sigma_l |C_m|^{-\frac{1}{5}} \tag{5}$$

where σ_l denotes the standard deviation of C_m for the l^{th} feature.

C. S-Divergence Between Two Clusters

The SD between two clusters is stated as follows:

Definition 2. Let C_m and C_n be two clusters. The M and N are the two probability mass functions (PMFs) of C_m and C_n respectively as defined in Eq. 4 with finite or countably infinite values in a discrete domain, $\mathcal{D}$. The SD between C_m and C_n is computed by Eq. 6.

$$D_S^2(M,N) = \log\left(\det\left(\frac{\phi_{|C_m|}(M) + \phi_{|C_m|}(N)}{2}\right)\right) - \frac{1}{2}\log\left(\det\left(\phi_{|C_m|}(M)\phi_{|C_m|}(N)\right)\right) \tag{6}$$

where we assume that M has C_m samples $M = \{x_1, x_2, ..., x_{|C_m|}\}$ and PMF of every uncertain object is converted into diagonal matrix using $\phi_{|C_m|}(\,)$ function as follows:

$$\phi_{|C_m|}(M) = \begin{bmatrix} M(x_1) & 0 & \dots & 0 \\ 0 & M(x_2) & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & M(x_{|C_m|}) \end{bmatrix}$$

$$\text{and } \phi_{|C_m|}(N) = \begin{bmatrix} N(x_1) & 0 & \dots & 0 \\ 0 & N(x_2) & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & N(x_{|C_m|}) \end{bmatrix}.$$

Sometimes, it is needed to smooth a PMF of a data object thus the probability values become non-negative in a domain since SD consists of a logarithmic function as shown in Eq. 6. Thus, Eq. 7 is employed for normalizing [1].

$$N'(x) = \frac{N(x) + \beta}{1 + \beta|\mathcal{D}|} \tag{7}$$

where β is a constant and the value of β lies between an interval [0, 1]. The $|\mathcal{D}|$ signifies the number of possible values in $\mathcal{D}$. Furthermore, the sum of integral of $N'(x)$ over the entire $\mathcal{D}$ is 1. Equation 8 is utilized to estimate error in smoothing.

$$|N'(x) - N(x)| = \left|\frac{1 - N(x)\beta}{\frac{1}{\beta} + |\mathcal{D}|}\right| \in \left[0, \frac{max\{1, |1 - \mathcal{D}|\}}{\frac{1}{\beta} + |\mathcal{D}|}\right] \tag{8}$$

The value of β is assigned to 0.001 in this work. The $\phi_{|C_m|}$ function is used to convert probability distributions to HPDM. The HPDM are manifolds, which are similar to non-positive curvature [31]. The HPDM cone does not come with a natural similarity function for a data object, although, it has computational and empirical advantages. Now, Eq. 6 is further simplified as follows:

$$D_S^2(M,N) = \log\left(\frac{1}{2^{|C_m|}} \times \det(\phi_{|C_m|}(M) + \phi_{|C_m|}(N))\right) + \log\left(\frac{1}{\sqrt{\det(\phi_{|C_m|}(M)\phi_{|C_m|}(N))}}\right) =$$

$$\log\left(\frac{1}{2^{|C_m|}} \times \frac{(M(x_1) + N(x_1))(M(x_2) + N(x_2)) \dots (M(x_{|C_m|}) + N(x_{|C_m|}))}{\sqrt{M(x_1)M(x_2) \dots M(x_{|C_m|})N(x_1)N(x_2) \dots g(x_{|C_m|})}}\right)$$

$$D_S^2(M,N) = \log\left(\frac{1}{2^{|C_m|}} \times \frac{(M(x_1) + N(x_1))}{\sqrt{(M(x_1)N(x_1))}} \times \frac{(M(x_2) + N(x_2))}{\sqrt{(M(x_2)N(x_2))}} \times \dots \times \frac{(M(x_{|C_m|}) + N(x_{|C_m|}))}{\sqrt{(M(x_{|C_m|})N(x_{|C_m|}))}}\right)$$

$$D_S^2(M,N) = \log\left(\frac{1}{2^{|C_m|}}\right) + \log\left(\left(\sqrt{\frac{M(x_1)}{N(x_1)}} + \sqrt{\frac{N(x_1)}{M(x_1)}}\right)\left(\sqrt{\frac{M(x_2)}{N(x_2)}} + \sqrt{\frac{N(x_2)}{M(x_2)}}\right) \dots \left(\sqrt{\frac{M(x_{|C_m|})}{N(x_{|C_m|})}} + \sqrt{\frac{N(x_{|C_m|})}{M(x_{|C_m|})}}\right)\right)$$

$$D_S^2(M,N) = \log\left(\frac{1}{2^{|C_m|}}\right) + \log\left(\sqrt{\frac{M(x_1)}{N(x_1)}}\right) + \log\left(\sqrt{\frac{M(x_2)}{N(x_2)}}\right) + \dots + \log\left(\sqrt{\frac{M(x_{|C_m|})}{g(x_{|C_m|})}}\right) + \log\left(1 + \frac{N(x_1)}{M(x_1)}\right) + \log\left(1 + \frac{N(x_2)}{M(x_2)}\right) + \dots + \log\left(1 + \frac{N(x_{|C_m|})}{M(x_{|C_m|})}\right)$$

$$D_S^2(M,N) = \log\left(\frac{1}{2^{|C_m|}}\right) + \sum_{x \in \mathcal{D}} \log\left(\sqrt{\frac{M(x)}{N(x)}}\left(1 + \frac{N(x)}{M(x)}\right)\right)$$

Finally, the SD between M and N is expressed as follows:

$$D_S^2(M,N) = \log\left(\frac{1}{2^{|C_m|}}\right) + \sum_{x \in \mathcal{D}} \log\left(\sqrt{\frac{|C_n|\prod_{l=1}^{d} h_l^N \sum_{\mathbf{c}_j \in C_m} \prod_{l=1}^{d} e^{-\frac{(x^l - c_{j,l})^2}{2h_l^{M^2}}}}{|C_m|\prod_{l=1}^{d} h_l^M \sum_{\mathbf{c}_j \in C_n} \prod_{l=1}^{d} e^{-\frac{(x^l - c_{j,l})^2}{2h_l^{N^2}}}}}\left(1 + \frac{|C_m|\prod_{l=1}^{d} h_l^M \sum_{\mathbf{c}_j \in C_n} \prod_{l=1}^{d} e^{-\frac{(x^l - c_{j,l})^2}{2h_l^{N^2}}}}{|C_n|\prod_{l=1}^{d} h_l^N \sum_{\mathbf{c}_j \in C_m} \prod_{l=1}^{d} e^{-\frac{(x^l - c_{j,l})^2}{2h_l^{M^2}}}}\right)\right)$$

D. The Proposed Internal CVI

The proposed internal CVI (PM) is based on CM and SM. So, the values of CM and SM need to be computed before calculating PM. The CM indicates the closeness or similarity of data objects in a cluster. Moreover, it is an average CM of all k clusters. The CM of every cluster, C_i, is calculated by Eq. 9. It is an average of aggregated squared S-distance (SD) of a cluster data object c_j to its center v_i.

$$CM = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{|C_i|}\sum_{\mathbf{c}_j \in C_i} D_{SD}\left(\mathbf{c}_j, \mathbf{v}_i\right) \tag{9}$$

where D_{SD} is the SD, which can be defined mathematically using Eq. 10.

Definition 3. define D_{SD}: $\Re_+^d \times \Re_+^d \to \Re_+ \cup \{0\}$ as

$$D_{SD}(\mathbf{c}_j, \mathbf{v}_i) = \sum_{l=1}^{d}[\log((\mathbf{c}_{j,l} + \mathbf{v}_{i,l})/2) - (\log(\mathbf{c}_{j,l}) + \log(\mathbf{v}_{i,l}))/2] \tag{10}$$

Equation 10 shows a point-to-point distance measure labeled as the SD that is motivated by the SD. It is defined in the open cone of PDM. Moreover, Eq. 10 shows that if two data objects with the same Euclidean distance are close to the origin, then data objects will have a larger SD compared to when they are far from the origin. This property can be applied to find the properties of clusters with varying sizes and densities. Furthermore, SD is neither an f-divergence nor a Bregman divergence and is invariant under the Hadamard product [45].

The CM ranges from 0 to ∞, where a low value is appropriate for a clustering configuration. The SM determines the magnitude of separation between clusters. The SD-based SM is calculated in this study by Eq. 11.

$$SM = \frac{2}{k(k-2)}\sum_{i=1}^{k}\sum_{j=i+1}^{k} D_S\left(M_i, M_j\right) \tag{11}$$

where M_i and M_j are the PMFs of clusters C_i and C_j respectively. The SM lies in the interval [0, ∞), where a high value implies good clustering. The PM is a ratio of the CM to the proposed SM, and it is estimated using Eq. 12.

$$PM = \frac{CM}{SM} \tag{12}$$

Good clustering is characterized by a low CM and a high SM of clusters. Therefore, a smaller value of PM is suitable for a clustering configuration. Sometimes, it is required to normalize the SM, and thus its value becomes non-zero in a domain since the zero value of SM will make an undefined value of the proposed index, PM, as shown in Eq. 12. Hence, Eq. 11 is further normalized.

$$SM = \frac{2}{k(k-2)}\sum_{i=1}^{k}\sum_{j=i+1}^{k} D_S\left(M_i, M_j\right) + \frac{\delta}{k^2} \tag{13}$$

where δ is a constant and the value of $\delta \to 0$, further estimated error in normalization is $\frac{\delta}{k^2}$ which is less significant in the possible range of SM. Normalized SM will be used throughout the paper to avoid an undefined value.

E. Complexity Analysis

The complexity associated with CM and SM is $O(nd)$ and $O(k^2dE)$ respectively, where E is the number of steps to estimate the SD between two clusters. The complexity of PM is represented by $O\left(nd + k^2dE\right)$ since $n \geq d$ and $n \geq k$ is considered in this study. Thus the complexity of the proposed CVI is linear.

IV. Experimental Results and Discussion

A laptop Intel(R) Core(TM) i7-2620M CPU@2.70GHz and 4-GB RAM running on Windows 10 having a 64-bits Python 3.6.5 compiler are considered for this study. All the work is carried out in Spyder 3.2.8's Python development environment.

A. Description of Databases

A total of 10 databases of two classes, namely synthetic and real-world are considered in this work to prove the effectiveness of the PM over some of the most popular existing internal CVIs. **Synthetic databases**: Three databases, namely Blobs, Varied Distributed data, and Anisotropically Distributed Data, are created in this study. The title of the databases, the total number of data objects in each database, the total number of features in each data object, and the number of clusters are noted in Table II. The Blobs database is produced by an isotropic Gaussian function with three classes having 1500 data objects or samples and two features. The varied distributed data is produced with varied variance in the data and has 1500 samples with 3 classes in 2D space, whereas Anisotropicly distributed database is generated by transforming the data, which is Anisotropically distributed or aligned on a specific axis. This database also has 1500 samples, three classes, and two features. **UCI and Kaggle repository databases**: Seven popular realistic databases, viz., Digits, Iris, Wine, Avila, Shuttle, Breast Cancer, and Letter Recognition, are adopted from the UCI repository [46], [47]. The short description of each of these UCI databases is also reported in Table II. All the databases are renamed as DB_i, where i varies from 1 to 10.

TABLE II. Datasets Characteristics

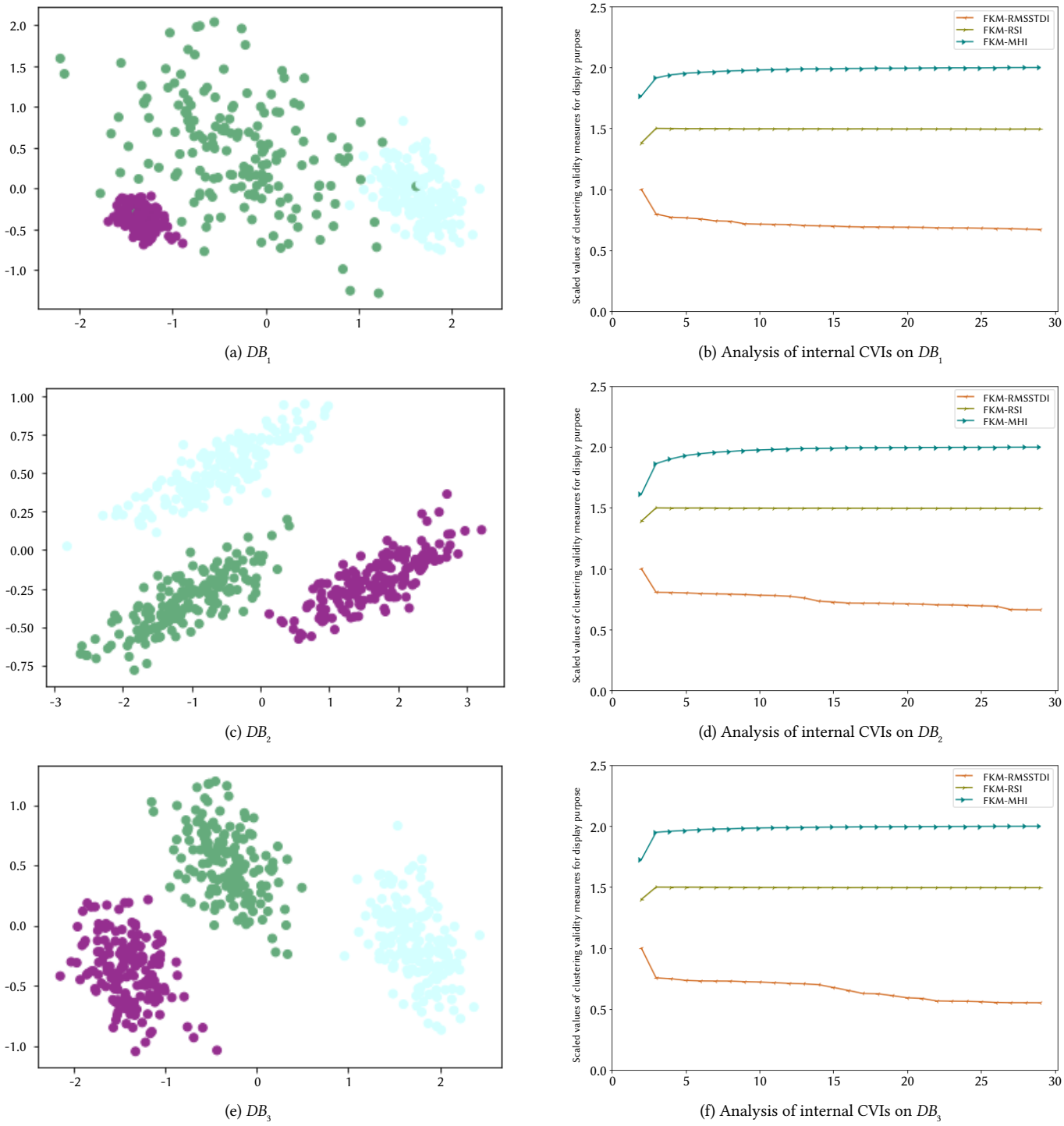
S. No.	Datasets	No of data objects	No of features	Clusters
1	Varied distributed data (DB1)	1500	2	3
2	Anisotropicly distributed data (DB2)	1500	2	3
3	Blobs (DB3)	1500	2	3
4	Breast cancer database (DB4)	569	30	2
5	Iris database (DB5)	150	4	3
6	Wine database (DB6)	178	13	3
7	Avila database (DB7)	10430	10	12
8	Digits database (DB8)	1797	64	10
9	Letter recognition database (DB9)	20000	16	26
10	Shuttle database (DB10)	43500	9	7

B. Results and Comparison

A couple of experiments are conducted to prove the effectiveness of PM over some of the existing internal CVIs in different scenarios, which are as follows:

1. The Impact of Monotonicity

The first experiment aims to study the monotonicity behavior of three internal CVIs, namely RMSSTDI, RSI, and MHI. Three synthetic databases, namely DB_1, DB_2, and DB_3 are considered, where clusters are well-separated. Fig. 2 (a), (c), and (e) plot the datasets DB_1, DB_2, and DB_3 along the x and y axes on a 2D plane, respectively. Here, fuzzy k-means (FKM) is applied to the three databases mentioned above, and the values of RMSSTDI, RSI, and MHI are computed, which are labeled as FKM-RMSSTDI, FKM-RSI, and FKM-MHI, respectively. Fig. 2 (b), (d), and (f) show the values of FKM-RMSSTDI, FKM-RSI, and FKM-MHI, respectively, that are obtained by varying the number of



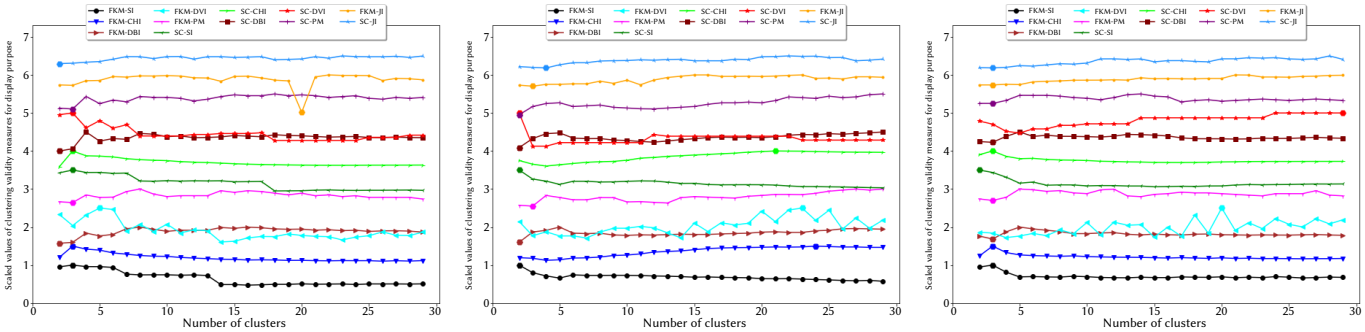
(a) DB_1 (b) Analysis of internal CVIs on DB_1

(c) DB_2 (d) Analysis of internal CVIs on DB_2

(e) DB_3 (f) Analysis of internal CVIs on DB_3

Fig. 2. DB_1, DB_2, and DB_3 are plotted on the plane, different classes are shown with different colors and result of internal CVIs on database in the right.

clusters, k, from 2 to 29 as inputs because the datasets discussed in Table II have an actual number of clusters in the range of 2 to 26. The other information on the results is not pertinent to this experiment. The vertical axis of curves or graphs in Fig. 2 is scaled for better visualization or analysis. When the value of k increases then value of numerator in $RMSSTDI = \frac{\sum_{i=1}^{k}\sum_{c_j \in c_i}||c_j - v_i||^2}{d(n-k)}$ will decrease. The value of $(n-k)$ is regarded as a constant because $k \ll n$. Therefore, RMSSTDI decreases with an increase in the k-value in Fig. 2 (b), (d), and (f). Further, RSI specifics a ratio of between clusters sum of squares to the total sum of squares. Hence, RSI increases as the value of k increases, as shown in Fig. 2 (b), (d), and (f). Similarly, MHI increases as the value of k increases, according to Fig. 2 (b), (d), and (f), because with an increase in k more pairs of distances are calculated. Furthermore, RMSSTDI is only based on CM, and RSI and MHI rely only on SM. According to the property of monotonicity, the curves of RMSSTDI, RSI, and MHI will be either downward or upward. It is quoted that the value of k is optimal at the "elbow" point, where a shift in the curve appears. Thus, the empirical results in Fig. 2 prove that the RMSSRDI, RSI, and MHI monotonically decrease or increase as the number of clusters, k, increases in the range from 2 to 29. However, the determination of a shift in the curve is rather a tedious and subjective task, thus the monotonicity is not discussed in the further sections.



(a) Analysis of internal CVIs on DB_1 (b) Analysis of internal CVIs on DB_2 (c) Analysis of internal CVIs on DB_3

Fig. 3. An analysis of internal CVIs on well-separated databases.

2. The Impact of Well-Separated Clusters

The aim of the 2nd experiment is to determine the optimal value of k for the databases, where well-separated clusters are present. The steps involved in estimating the optimal value of k for the best partitions using internal CVIs are as follows:

- Step 1: Initialize a clustering algorithm before applying it to a database.
- Step 2: A set of parameters of the algorithm is fixed in order to achieve clustering results.
- Step 3: Calculate the corresponding internal CVIs after clustering.
- Step 4: Select the optimal value of internal CVIs for best partitions.

Here, the values of six internal CVIs viz., SI, CHI, DBI, DVI, JI, and PM are computed after applying FKM and spectral clustering (SC) [48] on three databases, namely DB_1, DB_2, and DB_3 and results are reported in Fig. 3 (a), (b), and (c) respectively. The FKM-SI, FKM-CHI, FKM-DBI, FKM-DVI, FKM-JI, and FKM-PM specify the values of SI, CHI, DBI, DVI, JI, and PM after executing FKM while SC-SI, SC-CHI, SC-DBI, SC-DVI, SC-JI, and SC-PM are employed to represent the values of SI, CHI, DBI, DVI, JI, and PM after applying SC. Fig. 3 displays the values of FKM-SI, FKM-CHI, FKM-DBI, FKM-DVI, FKM-JI, FKM-PM, SC-SI, SC-CHI, SC-DBI, SC-DVI, SC-JI, and SC-PM that are obtained by varying the value of k in the range of 2 to 29. The optimal values of CVIs labeled by a hexagon marker in Fig. 3 specify either maximum or minimum values, which demonstrate the actual values of k in the databases. It is clear from Fig. 3 (a) that SC-PM, FKM-PM, FKM-JI, SC-JI, SC-DVI, SC-CHI, SC-SI, FKM-CHI, and FKM-SI determine the optimal value of k, which is the same as the exact number of clusters in DB_1. Moreover, the remaining CVIs produce values of k, which are closer to the actual number of clusters. It is also observed from Fig. 3 (b) that the FKM-PM and FKM-JI compute the optimal number of clusters, which are equal to the real number of clusters in DB_2. Furthermore, FKM-SI, FKM-DBI, SC-SI, SC-DBI, SC-DVI, SC-PM, FKM-JI, and SC-JI are also in proximity to the optimal clusters. On the other hand, the remaining CVIs are not near-optimal results. Fig. 3 (c) shows the results of DB_3 and that FKM-SI, FKM-CHI, FKM-DBI, FKM-PM, SC-DBI, SC-PM, FKM-JI, SC-JI, and SC-CHI achieve the optimal value for the clusters.

3. The Impact of Slightly Overlapped Clusters

The third experiment aims to decide the optimal value of k for the databases, namely DB_4, DB_5, and DB_6, where slightly overlapping clusters are present. However, principal component analysis is adopted in exploratory data analysis by transforming the data to a new coordinate system in the case of high-dimensional data and then plotting the first two principal components [49], [50]. The first two principal components of datasets DB_4, DB_5, and DB_6 are mapped on a 2D plane, which are displayed in Fig. 4 (a), (c), and (e), respectively. Here, slightly overlapping clusters are denoted by different colors. Again, the values of six internal CVIs, viz. SI, CHI, DBI, DVI, JI, and PM, are computed after applying FKM and SC on the three databases mentioned above, and the outcomes are noted in Fig. 4 (b), (d), and (f), respectively. Here, we run the clustering algorithms for different values of k in the range of 2 to 29. We can find out the exact values of k by considering the optimum values of the curves of the FKM-PM and SC-PM in most cases. Moreover, PM always helps to decide the exact value of k because of the use of non-linear similarity measures.

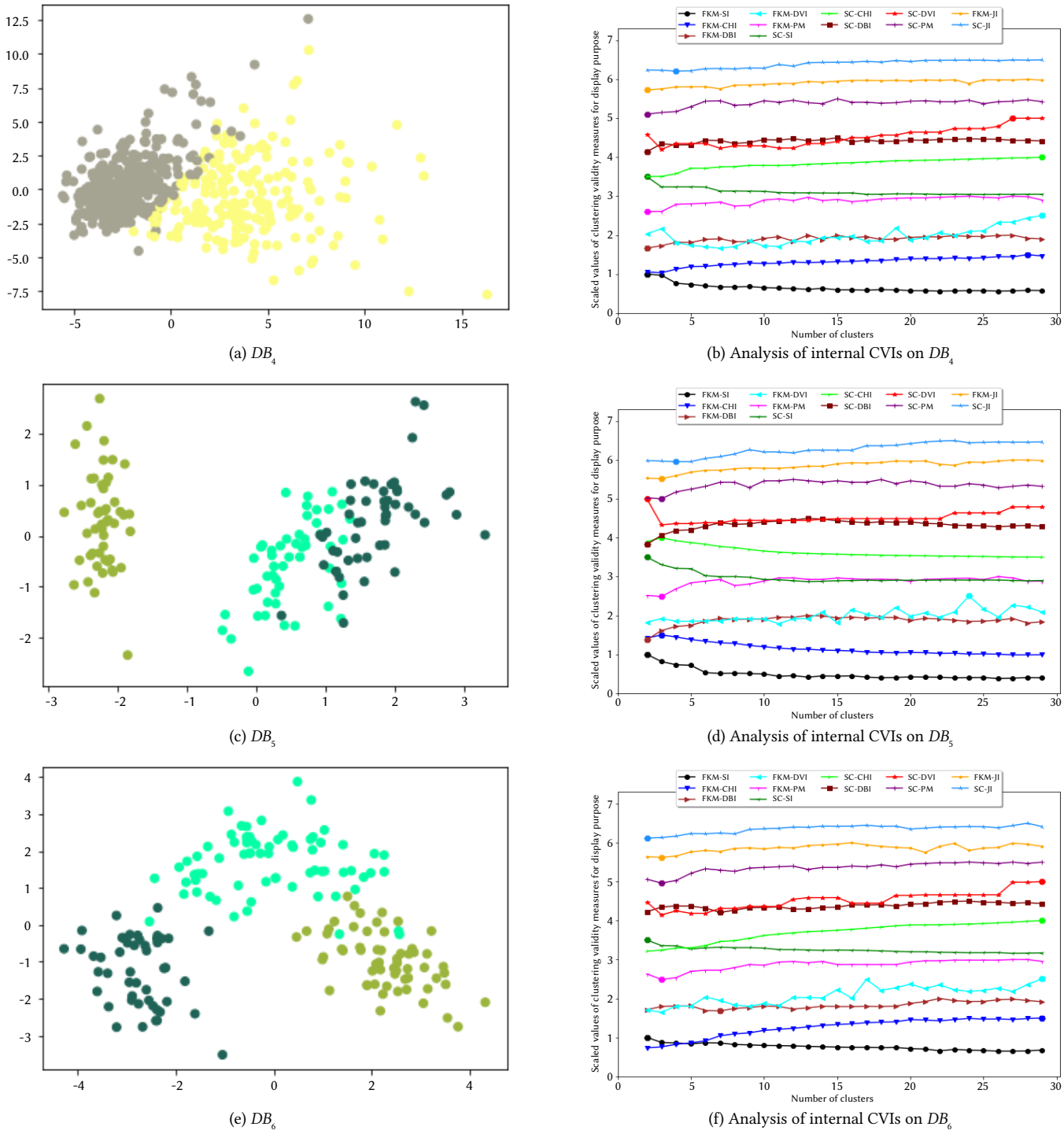
4. The Impact of Highly Overlapped Clusters

The focus of the fourth experiment is to estimate the optimal value of k for the databases, namely DB7, DB8, DB9, and DB10, where clusters are highly significant. The first two principal components of datasets DB_7, DB_8, DB_9, and DB_{10} are mapped on a 2D plane, which are displayed in Fig. 5 (a), (c), (e), and (g), respectively. Here, different colors are employed to represent clusters. Again, the values of six internal CVIs, viz. SI, CHI, DBI, DVI, JI, and PM, are calculated after applying FKM and SC on the four databases stated above, and the results are displayed in Fig. 5 (b), (d), (f), and (h), respectively. Here, both the clustering algorithms execute for different values of k in the range of 2 to 29. Focusing on the results, PM determines the optimal k for DB_7 and DB_8. But FKM-DBI, FKM-DVI, SC-SI, and SC-PM compute a value close to it. Furthermore, FKM-PM, SC-PM, and SC-DVI find the optimal k for DB_9, and FKM-DBI and FKM-DVI are not far from them. Finally, for DB_{10}, SC-PM and SC-DVI find the optimal k and FKM-DBI, FKM-DVI, FKM-PM, SC-DBI, and SC-JI compute a close value.

5. The Impact of Noise

The purpose of the 5th experiment is to determine how robust the proposed internal CVI named PM is against noisy features. First, noisy facets are included in the three well-separated databases, namely DB_1, DB_2, and DB_3. Here, a noisy feature is produced by considering uniform random distribution in the limit of the length and size similar to features of the original database. The number of features will be doubled in a database after adding noisy features. The impact of noisy features is then analyzed in this study. Databases are shown in Fig. 6 (a), (c), and (e). Again, the values of six internal CVIs, viz. SI, CHI, DBI, DVI, JI, and PM, are estimated after applying FKM and SC to the three noisy databases presented above, and the results are portrayed in Fig. 6 (b), (d), and (f), respectively. Here, both the clustering algorithms execute for different values of k in the range of 2 to 29. It is clear from Fig. 6 that DBI and DVI are affected by noise and face difficulty while determining the optimum value of k. Further, the curve of CHI is close to the optimal number of clusters in the case of a noisy DB2 database. On the other hand, the optimum values of SI, JI, and PM are closer to the exact values of k.

We can conclude from the five experiments conducted above that the proposed internal CVI named PM successfully ascertains the optimal number of clusters for most databases. On the other hand,



(a) DB_4 (b) Analysis of internal CVIs on DB_4

(c) DB_5 (d) Analysis of internal CVIs on DB_5

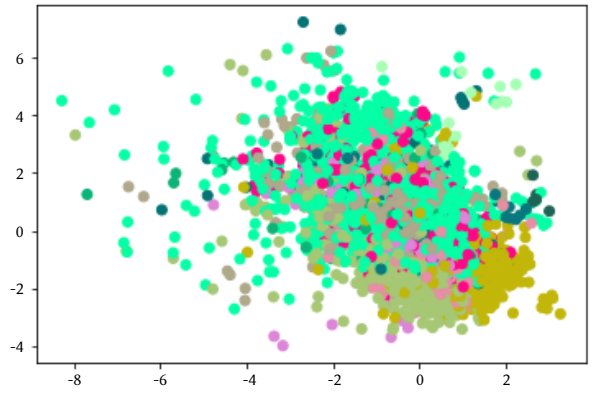
(e) DB_6 (f) Analysis of internal CVIs on DB_6

Fig. 4. In the left first two principal components of the DB_4, DB_5, and DB_6 are plotted on the plane, to display the first and second corresponding vectors of the data matrix along the axes, different classes are shown with different colors and result of internal CVIs on database in the right.

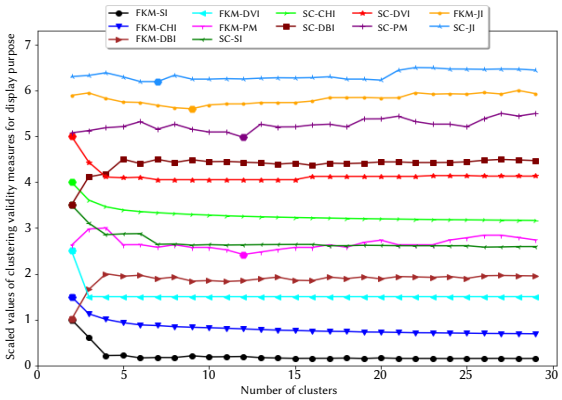
JI, SI, CHI, DBI, and DVI face difficulty while estimating the exact number of clusters due to various degrees of overlapping between clusters and noise in the databases.

6. The Comparative Analysis

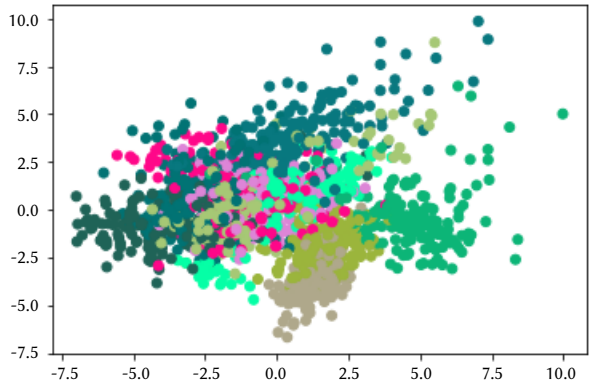
Finally, the PM is compared with five popular internal CVIs, namely SI, CHI, DBI, DVI, and JI, after applying four clustering algorithms, viz. FKM, SC, Density-based Spatial Clustering of Applications with Noise (DBSCAN), and Density Peak Clustering (DPC) [51] on the ten databases mentioned in Section IV A. Here, FKM and SC take the exact number of clusters as inputs, whereas DBSCAN and DPC compute the number of clusters automatically. The values of six internal CVIs, including the PM, are reported in Table III. The mean (μ) and standard deviation (σ) obtained by the four clustering algorithms of each CVI are also noted in the last column of Table III. The μ and σ of the PM are highlighted by bold characters. A smaller value of σ in percentage specifies well-separated and compact clusters. In other words, a smaller value of σ demonstrates that the clustering configuration is appropriate. It is clear from Table III that the PM consistently outperforms five considered internal CVIs on ten databases in different scenarios presented in Table IV. Therefore, PM can be a great choice while evaluating clustering results.



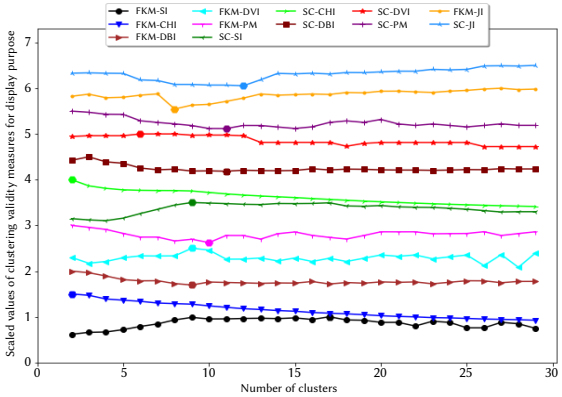
(a) DB_7



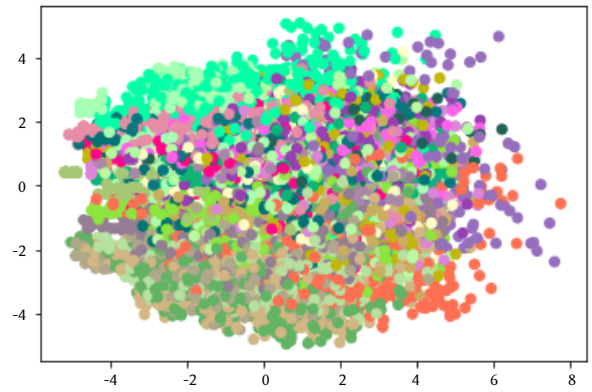
(b) Analysis of internal CVIs on DB_7



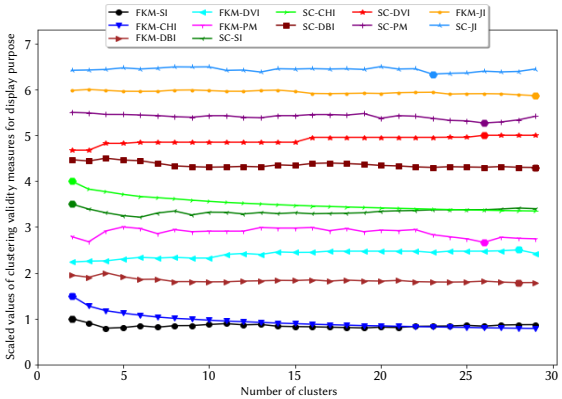
(c) DB_8



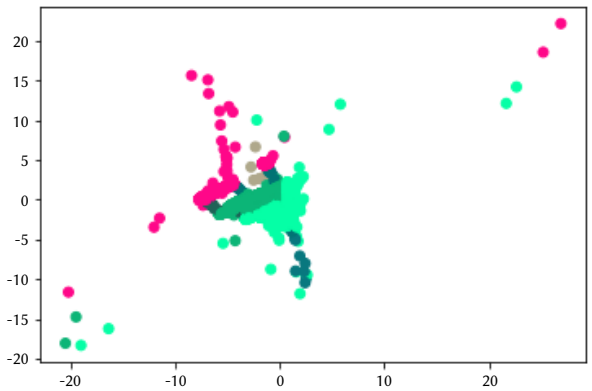
(d) Analysis of internal CVIs on DB_8



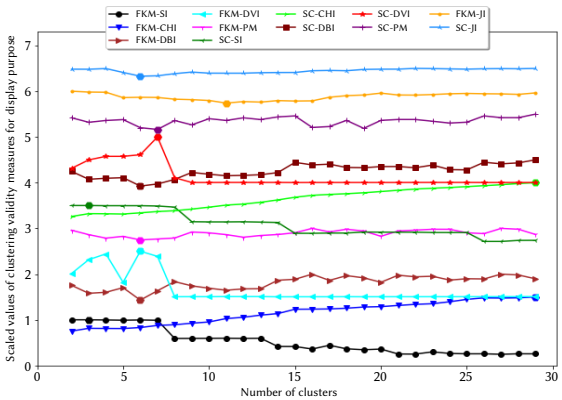
(e) DB_9



(f) Analysis of internal CVIs on DB_9



(e) DB_{10}

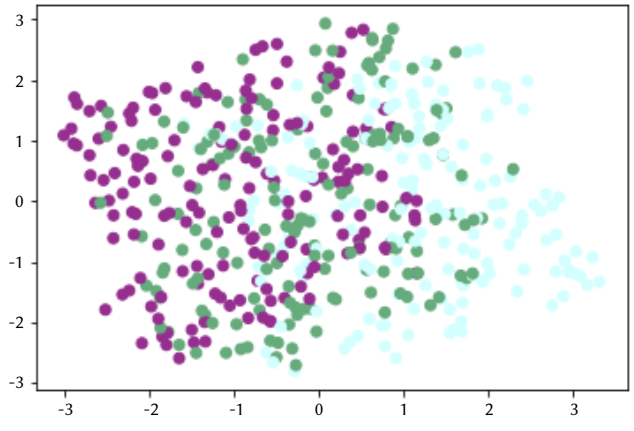


(f) Analysis of internal CVIs on DB_{10}

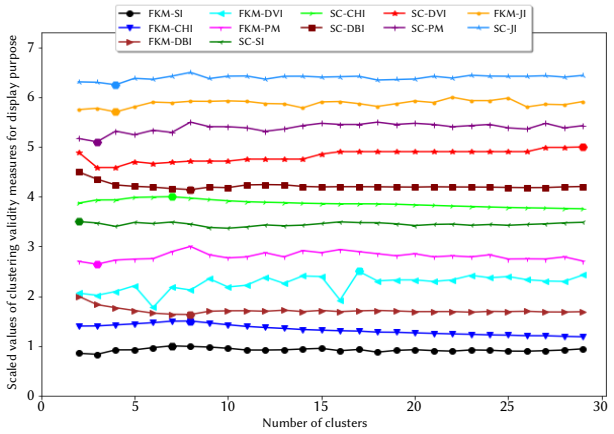
Fig. 5. In the left first two principal components of the DB7, DB8, DB9, and DB10 are plotted on the plane, to display the first and second corresponding vectors of the data matrix along the axes, different classes are shown with different colors and result of internal CVIs on database in the right.

TABLE III. Comparative Analysis of Internal CVIs Using Clustering Algorithms

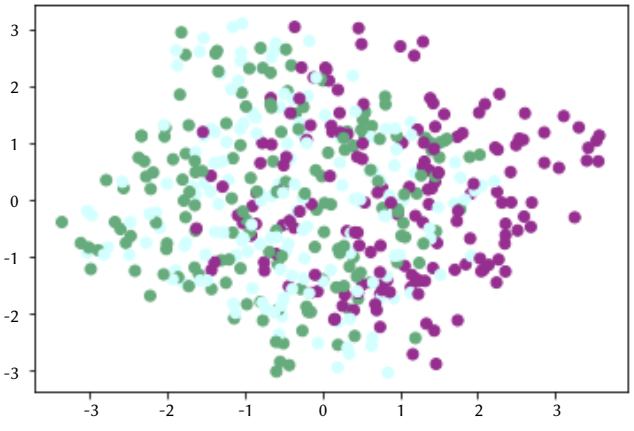
Dataset	CVI	FKM	SC	DBSCAN	DPC	$\mu \pm \sigma$ %
DB_1	SI	0.6468	0.61745	0.57755	0.62765	0.61736 ± 8.79207
	CHI	5451.4914	3810.65643	3792.7863	4756.62157	4452.88893 ± 18.04829
	DBI	0.56956	0.54442	0.8377	0.61122	0.64073 ± 20.94112
	DVI	0.00785	0.01372	0.02252	0.03266	0.01919 ± 56.37716
	JI	37.00884	39.56956	45.42931	34.47878	39.12162 ± 11.98999
	PM	**29.56956**	**30.33441**	**34.71643**	**28.57965**	30.80001 ± 4.72936
DB_2	SI	0.63551	0.48386	0.40725	0.50914	0.50894 ± 18.63688
	CHI	3883.88156	3302.55351	5465.96704	3803.73873	4114.03521 ± 22.78248
	DBI	0.49246	0.68642	0.71943	0.71491	0.65331 ± 16.56517
	DVI	0.00899	0.0069	0.00897	0.00376	0.00716 ± 34.47394
	JI	44.00376	41.71511	57.67286	43.40933	46.70027 ± 15.80087
	PM	**23.48942**	**30.68531**	**28.71825**	**22.72414**	26.40428 ± 14.78514
DB_3	SI	0.4863	0.42646	0.33207	0.46153	0.42659 ± 15.85287
	CHI	2011.98126	1601.38907	1413.97578	1481.83841	1627.29613 ± 16.46313
	DBI	0.73157	0.78999	0.84494	0.82526	0.79794 ± 6.23412
	DVI	0.00825	0.01911	0.01358	0.00881	0.01244 ± 40.60672
	JI	47.79319	43.01848	49.44894	44.48351	46.18603 ± 6.39381
	PM	**35.83244**	**39.79319**	**40.84494**	**37.81437**	38.57124 ± 5.74615
DB_4	SI	0.69726	0.50825	0.509	0.67526	0.59744 ± 17.23188
	CHI	1300.20823	1089.92944	1245.56763	1251.53446	1221.80994 ± 8.52372
	DBI	0.5044	0.62932	0.60906	0.55185	0.57366 ± 9.87326
	DVI	0.01731	0.00726	0.01246	0.02148	0.01463 ± 41.98165
	JI	60.5044	68.01731	74.07588	63.92288	66.6 ± 8.76054
	PM	**43.51121**	**49.63143**	**48.60891**	**41.56075**	45.82808 ± 7.46948
DB_5	SI	0.55282	0.55432	0.68674	0.68105	0.61873 ± 12.16705
	CHI	561.62776	558.05804	502.82156	513.92455	534.10798 ± 8.69226
	DBI	0.66197	0.64325	0.37927	0.39431	0.51970 ± 29.59093
	DVI	0.09881	0.12181	0.338	0.07651	0.15901± 76.31654
	JI	31.65626	32.11279	43.19802	32.38334	34.83760 ± 16.02200
	PM	**12.6709**	**14.65626**	**15.38275**	**13.40429**	14.02855 ± 5.63466
DB_6	SI	0.56448	0.57114	0.56067	0.56203	0.56458 ± 6.23471
	CHI	552.85171	561.81566	670.62599	708.08668	623.34501 ± 12.48562
	DBI	0.53573	0.53424	0.55357	0.54434	0.54197 ± 1.64643
	DVI	0.02237	0.01626	0.0374	0.03399	0.02751 ± 35.91714
	JI	48.01626	58.53573	51.64375	49.6353	51.95776 ± 8.91017
	PM	**27.53573**	**30.53424**	**31.55357**	**31.49413**	30.27942 ± 0.82341
DB_7	SI	0.1937	0.12995	0.1385	0.11951	0.14542 ± 22.77166
	CHI	5285.5617	4519.20875	4333.76871	4212.35646	4587.72391± 10.50701
	DBI	1.12112	1.29988	1.01121	0.8937	1.08148 ± 15.96829
	DVI	0.00182	0.00529	0.00197	0.00194	0.00276 ± 61.38810
	JI	24.12043	28.12995	35.10793	27.69293	28.76281 ± 15.97742
	PM	**8.12138**	**8.28694**	**7.57481**	**6.22372**	7.55171 ± 12.39664
DB_8	SI		0.1785	0.18289	0.17863	0.18066 ± 9.87675
	CHI	169.36261	161.20475	162.1034	171.6	166.07133 ± 3.12864
	DBI	1.9	1.88899	1.89937	1.84913	1.89023 ± 1.63817
	DVI	0.21933	0.26126	0.17384	0.19023	0.21117 ± 18.15176
	JI	42.87789	39.26069	49.92082	41.99865	43.51451 ± 10.43365
	PM	**15.92192**	**18.79859**	**18.90038**	**15.83872**	17.36490 ± 1.34074
DB_9	SI	0.1463	0.152	0.14713	0.139	0.14630 ± 6.85984
	CHI	142	496	146	7167	1376.25764 ± 6.35297
	DBI	1.6855	1.63312	1.64295	1.35005	1.57791 ± 9.73410
	DVI	0.04536	0.04307	0.04136	0.04036	0.04254 ± 5.14657
	JI	82.65005	96.04536	99.02207	94.98688	93.17609 ± 7.75123
	PM	**59.6855**	**66.62**	**60.63995**	**56.65005**	60.89963 ± 3.57705
DB_{10}	SI	0.97878	0.96967	0.97987	0.58508	0.87835 ± 22.26525
	CHI	15723.30982	14946.92039	12879.555	16331.2233	14970.25213 ± 10.05018
	DBI	0.34179	0.25082	0.3709	0.44054	0.35101 ± 22.39261
	DVI	0.13045	0.24059	0.04701	0.09064	0.12717 ± 65.21501
	JI	61.68367	58.31793	64.39526	63.44635	61.96080 ± 4.31864
	PM	**40.33581**	**39.25111**	**43.36991**	**41.43915**	41.09900 ± 4.27706



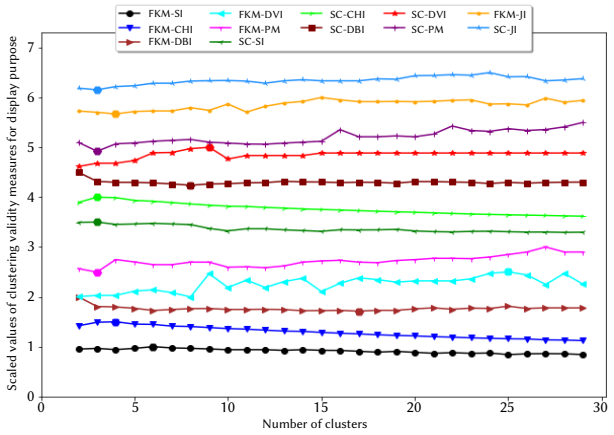
(a) Noisy-DB_1



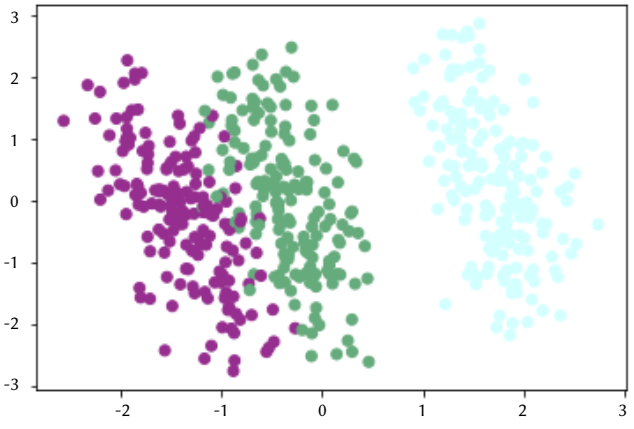
(b) Analysis of internal CVIs on Noisy-DB_1



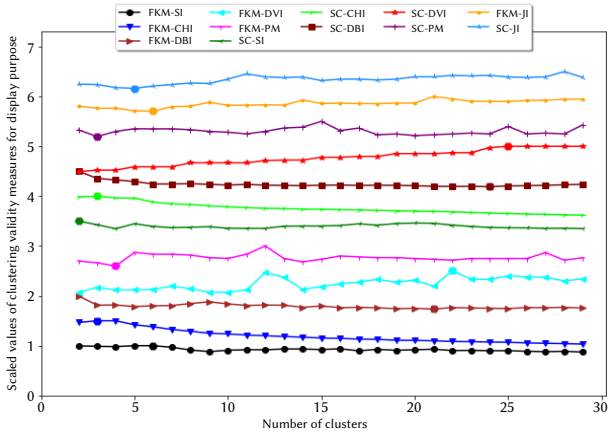
(c) Noisy-DB_2



(d) Analysis of internal CVIs on Noisy-DB_2



(e) Noisy-DB_3



(f) Analysis of internal CVIs on Noisy-DB_3

Fig. 6. In the left noisy-databases are plotted on the plane, different classes are shown with different colors and result of internal CVIs on noisy-database in the right.

TABLE IV. The Overall Review of Some Internal CVIs

Index	Well-separated	Slightly-separated	Highly-overlapped	Noise
SI	G	G	X	A
CHI	G	G	X	A
DVI	G	A	X	X
DBI	A	G	X	A
JI	G	G	A	A
PM	G	G	G	G

V. Conclusion

Internal CVIs are employed frequently in clustering to measure the goodness of the clustering algorithms without taking any external inputs. Most of the existing internal CVIs depend on CM and the geometric distance-based SM when computing the distance between cluster centers. The previous studies showed that such CVIs are not capable of producing accurate results, especially when the clusters of a database are highly overlapping. As a remedy, we introduce a new internal CVI, PM, using a modified CM and an updated SM based on the notion of SD. Moreover, SD is defined on the cone of HPDM and is

shown to have experimental and computational advantages over the other approaches in many applications. On the other hand, SD is a point-to-point distance measure that is motivated by the definition of SD. It is defined in the open cone of PDM. Initially, clusters of a database are modeled using density functions by applying a non-parametric kernel density estimation method. The PM is defined as the ratio of the modified CM to the updated SM. A smaller value of the PM indicates that the clustering configuration is appropriate. Empirical results illustrate that the PM is proficient in determining the exact number of clusters and the best partition for several superficial and realistic databases, including the database with arbitrary cluster shapes. The proposed internal CVI faces difficulty in ascertaining the optimal number of clusters when noisy features are included in a few databases. In addition, the proposed internal CVI works efficiently for databases having only numerical attributes. The latter two aspects deserve further study. In future work, SD may be explored to develop an external CVI.

Acknowledgment

This work is partially supported by the project "Smart Solutions in Ubiquitous Computing Environments", Grant Agency of Excellence (under ID: UHKFIM-GE-2023), University of Hradec Kralove, Faculty of Informatics and Management, Czech Republic.

Conflict of interest

The authors declare no conflict of interest.

References

[1] K. K. Sharma, A. Seal, "Modeling uncertain data using monte carlo integration method for clustering," *Expert Systems with Applications*, vol. 137, pp. 100-116, 2019.

[2] K. K. Sharma, A. Seal, "Clustering analysis using an adaptive fused distance," *Engineering Applications of Artificial Intelligence*, vol. 96, p. 103928, 2020.

[3] A. Seal, A. Karlekar, O. Krejcar, E. Herrera-Viedma, "Performance and convergence analysis of modified c-means using jeffreys-divergence for clustering," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 2, pp. 141-149, 2021.

[4] M. Martín Merino, A. J. López Rivero, V. Alonso, M. Vallejo, A. Ferreras, "A clustering algorithm based on an ensemble of dissimilarities: An application in the bioinformatics domain," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 6, pp. 6-13, 2022.

[5] E. Asensio, A. Almeida, A. Galiano, J.-M. Martín- Álvarez, "Using customer knowledge surveys to explain sales of postgraduate programs: A machine learning approach," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 3, pp. 96-102, 2022.

[6] F. A. Ozbay, B. Alatas, "Fake news detection within online social media using supervised artificial intelligence algorithms," *Physica A: Statistical Mechanics and its Applications*, vol. 540, p. 123174, 2020.

[7] B. K. Dedeturk, B. Akay, "Spam filtering using a logistic regression model trained by an artificial bee colony algorithm," *Applied Soft Computing*, vol. 91, p. 106229, 2020.

[8] S. Munusamy, P. Murugesan, "Modified dynamic fuzzy c-means clustering algorithm–application in dynamic customer segmentation," *Applied Intelligence*, pp. 1–21, 2020.

[9] I.-C. Wu, H.-K. Yu, "Sequential analysis and clustering to investigate users' online shopping behaviors based on need-states," *Information Processing & Management*, vol. 57, no. 6, p. 102323, 2020.

[10] A. Sivanathan, H. H. Gharakheili, V. Sivaraman, "Detecting behavioral change of iot devices using clustering-based network traffic modeling," *IEEE Internet of Things Journal*, vol. 7, no. 8, pp. 7295–7309, 2020.

[11] A. Das, J. Nayak, B. Naik, U. Ghosh, "Generation of overlapping clusters constructing suitable graph for crime report analysis," *Future Generation Computer Systems*, vol. 118, pp. 339–357, 2021.

[12] A. K. Tripathi, K. Sharma, M. Bala, A. Kumar, V. G. Menon, A. K. Bashir, "A parallel military-dog-based algorithm for clustering big data in cognitive industrial internet of things," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 3, pp. 2134–2142, 2021, doi: 10.1109/TII.2020.2995680.

[13] M. Landauer, F. Skopik, M. Wurzenberger, A. Rauber, "System log clustering approaches for cyber security applications: A survey," *Computers & Security*, vol. 92, p. 101739, 2020.

[14] A. K. Abasi, A. T. Khader, M. A. Al-Betar, S. Naim, S. N. Makhadmeh, Z. A. A. Alyasseri, "Link-based multi- verse optimizer for text documents clustering," *Applied Soft Computing*, vol. 87, p. 106002, 2020.

[15] S. Lin, K. Schorpp, I. Rothenaigner, K. Hadian, "Image- based high-content screening in drug discovery," *Drug discovery today*, 2020.

[16] A. Belhadi, Y. Djenouri, J. C.-W. Lin, C. Zhang, A. Cano, "Exploring pattern mining algorithms for hashtag retrieval problem," *IEEE Access*, vol. 8, pp. 10569– 10583, 2020.

[17] A. Karlekar, A. Seal, O. Krejcar, C. Gonzalo-Martin, "Fuzzy k-means using non-linear s-distance," *IEEE Access*, vol. 7, pp. 55121–55131, 2019.

[18] A. Seal, A. Karlekar, O. Krejcar, C. Gonzalo-Martin, "Fuzzy c-means clustering using jeffreys-divergence based similarity measure," *Applied Soft Computing*, vol. 88, p. 106016, 2020.

[19] K. K. Sharma, A. Seal, "Spectral embedded generalized mean based k-nearest neighbors clustering with s-distance," *Expert Systems with Applications*, vol. 169, p. 114326, 2021.

[20] K. K. Sharma, A. Seal, A. Yazidi, A. Selamat, O. Krejcar, "Clustering uncertain data objects using jeffreys- divergence and maximum bipartite matching based similarity measure," *IEEE Access*, vol. 9, pp. 79505-79519, 2021.

[21] A. Seal, E. Herrera Viedma, *et al.*, "Performance and convergence analysis of modified c-means using jeffreys-divergence for clustering," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 2, pp. 141-149, 2021.

[22] K. K. Sharma, A. Seal, A. Yazidi, O. Krejcar, "A new adaptive mixture distance-based improved density peaks clustering for gearbox fault diagnosis," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–16, 2022, doi: 10.1109/TIM.2022.3216366.

[23] T. Ullmann, C. Hennig, A.-L. Boulesteix, "Validation of cluster analysis results on validation data: A systematic framework," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, p. e1444, 2022.

[24] B. Tavakkol, J. Choi, M. K. Jeong, S. L. Albin, "Object- based cluster validation with densities," *Pattern Recognition*, vol. 121, p. 108223, 2022.

[25] K. K. Sharma, A. Seal, "Multi-view spectral clustering for uncertain objects," *Information Sciences*, vol. 547, pp. 723-745, 2020.

[26] K. K. Sharma, A. Seal, "Outlier-robust multi-view clustering for uncertain data," *Knowledge-Based Systems*, vol. 211, p. 106567, 2021.

[27] K. K. Sharma, A. Seal, E. Herrera-Viedma, O. Krejcar, "An enhanced spectral clustering algorithm with s- distance," *Symmetry*, vol. 13, no. 4, p. 596, 2021.

[28] B. Liang, J. Cai, H. Yang, "A new cell group clustering algorithm based on validation & correction mechanism," *Expert Systems with Applications*, vol. 193, p. 116410, 2022.

[29] H. Cui, M. Xie, Y. Cai, X. Huang, Y. Liu, "Cluster validity index for adaptive clustering algorithms," *IET Communications*, vol. 8, no. 13, pp. 2256–2263, 2014.

[30] B. Tang, S. Kay, H. He, "Toward optimal feature selection in naive bayes for text categorization," *IEEE transactions on knowledge and data engineering*, vol. 28, no. 9, pp. 2508–2521, 2016.

[31] S. Sra, "Positive definite matrices and the s- divergence," *Proceedings of the American Mathematical Society*, vol. 144, no. 7, pp. 2787–2797, 2016.

[32] A. Fahad, N. Alshatri, Z. Tari, A. Alamri, I. Khalil, A. Y. Zomaya, S. Foufou, A. Bouras, "A survey of clustering algorithms for big data: Taxonomy and empirical analysis," *IEEE transactions on emerging topics in computing*, vol. 2, no. 3, pp. 267–279, 2014.

[33] S. Sharma, "Applied multivariate techniques, jhonn wiley & sons inc.; 116, new york," *Lewis-Beck vd*, vol. 1994, pp. 112–113, 1996.

[34] L. Hubert, P. Arabie, "Comparing partitions," *journal of classification*, vol. 2, no. 1, pp. 193–218, 1985.

[35] T. Caliński, J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics-theory and Methods*, vol. 3, no. 1, pp. 1–27, 1974.

[36] J. C. Dunn, "Well-separated clusters and optimal fuzzy partitions," *Journal of cybernetics*, vol. 4, no. 1, pp. 95– 104, 1974.
[37] D. L. Davies, D. W. Bouldin, "A cluster separation measure," *IEEE transactions on pattern analysis and machine intelligence*, no. 2, pp. 224–227, 1979.
[38] A. B. Said, R. Hadjidj, S. Foufou, "Cluster validity index based on jeffrey divergence," *Pattern Analysis and Applications*, vol. 20, no. 1, pp. 21–31, 2017.
[39] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.
[40] U. Maulik, S. Bandyopadhyay, "Performance evaluation of some clustering algorithms and validity indices," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 12, pp. 1650–1654, 2002.
[41] X. L. Xie, G. Beni, "A validity measure for fuzzy clustering," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 13, no. 8, pp. 841–847, 1991.
[42] M. Bouguessa, S. Wang, H. Sun, "An objective approach to cluster validation," *Pattern Recognition Letters*, vol. 27, no. 13, pp. 1419–1430, 2006.
[43] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Perez, I. Perona, "An extensive comparative study of cluster validity indices," Pattern Recongnition, vol. 46, no. 1, pp. 243-256, 2013.
[44] S. Sra, R. Hosseini, "Conic geometric optimization on the manifold of positive definite matrices," *SIAM Journal on Optimization*, vol. 25, no. 1, pp. 713–739, 2015.
[45] S. Chakraborty, S. Das, "k- means clustering with a new divergence-based distance metric: Convergence and performance analysis," *Pattern Recognition Letters*, vol. 100, pp. 67–73, 2017.
[46] C. De Stefano, M. Maniaci, F. Fontanella, A. S. di Freca, "Reliable writer identification in medieval manuscripts through page layout features: The "avila" bible case," *Engineering Applications of Artificial Intelligence*, vol. 72, pp. 99–110, 2018.
[47] D. Dheeru, E. Karra Taniskidou, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml.
[48] S. Affeldt, L. Labiod, M. Nadif, "Spectral clustering via ensemble deep autoencoder learning (sc-edae)," *Pattern Recognition*, vol. 108, p. 107522, 2020.
[49] S. Wold, K. Esbensen, P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987.
[50] C. Ding, X. He, "K-means clustering via principal component analysis," in *Proceedings of the twenty-first international conference on Machine learning*, 2004, p. 29.
[51] L. Bai, X. Cheng, J. Liang, H. Shen, Y. Guo, "Fast density clustering strategies based on the k-means algorithm," *Pattern Recognition*, vol. 71, pp. 375–386, 2017.

Krishna Kumar Sharma

He received PhD from the Computer Science and Engineering Department, PDPM Indian Institute of Information Technology, Design and Manufacturing Jabalpur, Madhya Pradesh, India in 2021 and has received the M.Tech.(Information Technology) degree from IIIT Allahabad, Uttar Pradesh, India, in 2011. He is currently an Assistant Professor with the Computer Science and Informatics Department, University of Kota, Kota, Rajasthan, India. His current research interest includes pattern recognition.

Ayan Seal

He received a Ph.D. in engineering from Jadavpur University, West Bengal, India, in 2014. He visited the Universidad Politecnica de Madrid, Spain as a visiting research scholar. He is currently an Assistant Professor with the Computer Science and Engineering Department, PDPM Indian Institute of Information Technology, Design and Manufacturing Jabalpur, Jabalpur, Madhya Pradesh, 482005, India. He is the recipient of several awards. He is at the top %2 scientists according to Stanford University, 2022. Dr. Seal has been granted sponsored projects by the Govt. of India funding agencies. He has authored or co-authored several journals, conferences, and book chapters on the applications of computer vision. He is on the editorial board of several journals. His current research interests include computer vision, machine learning, deep learning, and brain-computer interface.

Anis Yazidi

He received the M.Sc. and Ph.D. degrees from the University of Agder, Grimstad, Norway, in 2008 and 2012, respectively. He was a Researcher with Teknova AS, Grimstad, Norway. From 2014 to 2019, he was an Associate Professor with the Department of Computer Science, Oslo Metropolitan University, Oslo, Norway, where he is currently a Full Professor, leading the research group in applied artificial intelligence. He is also Professor II with the Norwegian University of Science and Technology (NTNU), Trondheim, Norway. His current research interests include machine learning, learning automata, stochastic optimization, and autonomous computing.

Ondrej Krejcar

He is a full professor in systems engineering and informatics at the University of Hradec Kralove, Faculty of Informatics and Management, Center for Basic and Applied Research, Czech Republic; and Research Fellow at Malaysia-Japan International Institute of Technology, University Technology Malaysia, Kuala Lumpur, Malaysia. In 2008 he received his Ph.D. title in technical cybernetics at Technical University of Ostrava, Czech Republic. He is currently a vice-rector for science and creative activities of the University of Hradec Kralove from June 2020. At present, he is also a director of the Center for Basic and Applied Research at the University of Hradec Kralove. In years 2016-2020 he was vice-dean for science and research at Faculty of Informatics and Management, UHK. His h-index is 23 according Web of Science, with more than 2500 citations received in the Web of Science, where more than 150 IF journal articles is indexed in JCR index (h-index 27 at SCOPUS with more than 3200 citations). In 2018, he was the 14th top peer reviewer in Multidisciplinary in the World according to Publons and a Top Reviewer in the Global Peer Review Awards 2019 by Publons. Currently, he is on the editorial board of the MDPI Sensors IF journal (Q1/Q2 at JCR), and several other ESCI indexed journals. He is a Vice-leader and Management Committee member at WG4 at project COST CA17136, since 2018. He has also been a Management Committee member substitute at project COST CA16226 since 2017. Since 2019, he has been Chairman of the Program Committee of the KAPPA Program, Technological Agency of the Czech Republic as a regulator of the EEA/Norwegian Financial Mechanism in the Czech Republic (2019-2024). Since 2020, he has been Chairman of the Panel 1 (Computer, Physical and Chemical Sciences) of the ZETA Program, Technological Agency of the Czech Republic. Since 2014 until 2019, he has been Deputy Chairman of the Panel 7 (Processing Industry, Robotics, and Electrical Engineering) of the Epsilon Program, Technological Agency of the Czech Republic. At the University of Hradec Kralove, he is a guarantee of the doctoral study program in Applied Informatics, where he is focusing on lecturing on Smart Approaches to the Development of Information Systems and Applications in Ubiquitous Computing Environments. His research interests include Technical Cybernetics, Ubiquitous Computing, Control Systems, Smart Sensors, Wireless Technology, Biomedicine, Image Segmentation and Recognition, Biometrics, Biotelemetric System Architecture (portable device architecture, wireless biosensors), development of applications for mobile / remote devices with use of remote or embedded biomedical sensors.