# Clutching of Clustering Validation Criteria

Urvashi Soni* and Sunita Dwivedi

Computer Department, Makhanlal Chaturvedi National University of Journalism and Communication, India
Email: urvashikushsoni2016@gmail.co (U.S.); ddwivedi2001@gmail.com (S.D.)
*Corresponding author
Manuscript received December 12, 2023; revised January 12, 2024; accepted January 30, 2024; published February 29, 2024

***Abstract*—The success of clustering depends critically on a number of key concerns, one of which is clustering validation. In general, there are three types of clustering validation criteria: relative clustering validation, internal clustering validation, and external clustering validation. This paper focuses on the clustering validation criteria and provides a thorough analysis of the most popular clustering validation for crisp clustering. We investigate the validation properties over the five conventional clustering. According to experiment results, Silhouette is the validation measure that performs well in all five areas whereas other measures have some limits in various application.**

***Keywords*—voting, consensus clustering, ensemble generation, co-occurrence matrices**

## I. Introduction

The task of grouping a set of items into clusters such that things inside the same cluster are similar and separate clusters are distinct is known as clustering, which is one of the most significant unsupervised learning. In several disciplines, including image analysis and bioinformatics, clustering is often used. It is essential to figure out a method to validate the validity of partitions following clustering because this is an unsupervised learning. Otherwise, utilising various clustering results would be challenging. One of the important requirements necessary for the success of clustering applications is clustering validation [1], which assesses the goodness of clustering outcomes [2]. The three basic types of clustering validation are External Clustering Validation, Relative Clustering Validation and Internal Clustering Validation. Entropy is an illustration of an external validation metric; it assesses the "purity" of clusters based on the class labels [3].

Internal validation measures use information found in the data, as opposed to external validation measures, which also use information not found in the data. Without taking into account outside data, internal measurements assess a clustering structure's [4]. External validation measures are mostly used to select the best clustering technique for a given data set because they are aware of the "actual" cluster number beforehand. On the other hand, internal validation measures can be utilised without any additional information to select the optimum clustering technique and the appropriate cluster size. In actuality, many application scenarios lack access to external data like class labels. Therefore, internal validation measures are the choice for cluster validation when there is no external information accessible.

External criteria suggest that the outcomes of a clustering algorithm are assessed in accordance with a predetermined structure that is put on a dataset to reflect the dataset's clustering structure. In other words, external criteria are grounded in the dataset's a priori knowledge or ground truth. In this paper, we will present four index: the Jaccard Index (JI), the Rand Index (RI), and its derivative, the adjusted Rand Index (ARI), (NMI). Internal criteria, as opposed to a priori knowledge, evaluate the clustering algorithms in terms of the internal structures of the datasets themselves. Re-sampling is the frequently used by this class of algorithms. This group of algorithms may provide accurate estimates of the number of clusters present in a dataset as well as reliable feedback on the performance of clustering techniques.

According to the relative correlation between compactness and separation, relative criteria assess the clustering partitions. The index values are used to evaluate clustering partitions rather than clustering techniques. The various subclasses of relative criteria include model-based index, fuzzy validity index, and crisp validity index. Calinski-Harabasz (CH), Dunn's (DI), Davies-Bouldin (DB), I index (II), Silhouette (SIL), Object-Based Validation (OBV-LDA), the Geometrical Index (GI), and the validity index are some of the crisp validity index. DB is the counterpart of XB in crisp clustering.

Numerous clustering validation measures, including *CH*, *I*, *DB*, *SD*, and *SIL*, have been put out in the paper for crisp clustering. However, a number of data properties can have an impact on the current measures. If minimum or maximum pairwise distances are utilised in the measure, for instance, data noise can significantly affect how well the validation measure performs. It is yet unknown how well-performing existing measures are in various scenarios. As a result, Table 1 presents a thorough analysis of 10 frequently used validation metrics. We examine the five different features of their validation properties: monotonicity, noise, density, subclusters, and skewed distributions. We create fictitious data for experiments for each aspect. These artificial data accurately reflect the qualities. *SIL* is the only validation measure that performs well in all five characteristics, according to the experiment results, although other measures have distinct limits in various application circumstances, particularly in terms of noise and sub clusters.

## II. Clustering Validation Measures

We discuss several fundamental validation concepts in this part, along with a group of 10 popular validation index. Validation methods are frequently based on the two criteria listed below [4, 5], as the purpose of clustering is to make things within the same cluster similar and those in different clusters distinguishable.

Compactness: It determines how tightly a cluster's objects are connected to one another. Based on variance, a set of measurements assesses cluster compactness. Better

compactness is indicated by lower variance. There are also many distance-based measures that are used to gauge how compact a cluster is, including maximum or average pairwise distances and maximum or standard core distances (see Table 1).

Table 1. Internal clustering validation measures

| S.no. | Measure | Notation | Optimal value | Definition |
|---|---|---|---|---|
| **1** | Calinski-Harabasz index | *CH* | Max | $\sum_i \boldsymbol{n_i d^2(c_i, c)/(NC-1)/\sum_i \sum_{x\in C_i} d^2\,(x, c_i)/(n-NC)}$ |
| **2** | Davies-Bouldin index | *DB* | Min | $\frac{1}{NC}\sum_i \boldsymbol{max}_{j,j\neq i}\left\{\left[\frac{1}{n_i}\sum_{x\in C_i} \boldsymbol{d}(x, c_i) + \frac{1}{n_j}\sum_{x\in C_j} \boldsymbol{d}(x, c_j)\right]/\boldsymbol{d}(c_i, c_j)\right\}$ |
| **3** | Dunn's index | *D* | Max | $\boldsymbol{min_i}\left\{\boldsymbol{min_j}\left(\boldsymbol{min}_{x\in C_i, y\in C_j}\boldsymbol{d}(x,y)/\boldsymbol{max_k}\left\{\boldsymbol{max}_{x,y\in C_k}\boldsymbol{d}(x,y)\right\}\right)\right\}$ |
| **4** | Modified Hubert Γ statistic | Γ | Elbow | $\frac{2}{n(n-1)}\sum_{x\in D}\sum_{y\in D}\boldsymbol{d}(x,y)\,\boldsymbol{d}_{x\in C_i, y\in C_j}(c_i, c_j)$ |
| **5** | Root-mean-square std dev | *RMSSTD* | Elbow | $\left\{\sum_i \sum_{x\in C_i}\|x-c_i\|^2/[P\ \sum_i(n_i-1)]\right\}^{\frac{1}{2}}$ |
| **6** | R-squared | *RS* | Elbow | $\left(\sum_{x\in D}\|x-c\|^2 - \sum_i\sum_{x\in C_i}\|x-c_i\|^2\right)/\sum_{x\in D}\|x-c\|^2$ |
| **7** | SD validity index | *SD* | Min | $\boldsymbol{Dis(NC_{max})Scat(NC) + Dis(NC)}$<br>$\boldsymbol{Scat(NC)} = \frac{1}{NC}\sum_i \|\sigma(C_i)\|/\|\sigma(D)\|\,,\boldsymbol{Dis(NC)}$<br>$= \frac{\boldsymbol{max}_{i,j}\,\boldsymbol{d}(c_i,c_j)}{\boldsymbol{min}_{i,j}\,\boldsymbol{d}(c_i,c_j)}\sum_i\left(\sum_j \boldsymbol{d}(c_i,c_j)\right)^{-1}$ |
| **8** | Silhouette index | *SIL* | Max | $\frac{1}{NC}\sum_i\left\{\frac{1}{n_i}\sum_{x\in C_i}\frac{b(x)-a(x)}{max[b(x),a(x)]}\right\}$<br>$\boldsymbol{a(x)} = \frac{1}{n_i-1}\sum_{y\in C_i, y\neq x}\boldsymbol{d}(x,y), \boldsymbol{b(x)} = \boldsymbol{min}_{j,j\neq i}\left[\frac{1}{n_j}\sum_{y\in C_j}\boldsymbol{d}(x,y)\right]$ |
| **9** | Xie-Beni index | *XB* | Min | $\left[\sum_i\sum_{x\in C_i}\boldsymbol{d^2}(x,c_i)\right]/\left[n.\boldsymbol{min}_{i,j\neq i}\boldsymbol{d^2}(c_i,c_j)\right]$ |
| **10** | *I* index | *I* | Max | $\left(\frac{1}{NC}.\frac{\sum_{x\in D}\boldsymbol{d}(x,c)}{\sum_i\sum_{x\in C_i}\boldsymbol{d}(x,c_i)}.\boldsymbol{max}_{i,j}\,\boldsymbol{d}(c_i,c_j)\right)^{\boldsymbol{p}}$ |

*D*: data set; *n*: number of objects in *D*; *c*: centre of *D*; *P*: attributes number of *D*; *NC*: number of clusters; $C_i$: the i[th] cluster; $n_i$: number of objects in $C_i$; $c_i$: centre of $C_i$; $\sigma(C_i)$: variance vector of $C_i$; $d(x, y)$: distance between x and y; $\|\boldsymbol{X_i}\| = (\boldsymbol{X_i^T}.\boldsymbol{X_i})^{\frac{1}{2}}$

Separation: It evaluates a cluster's degree of differentiation or separation from other clusters. Examples of measures of separation include the pairwise distances between cluster centers or the pairwise minimum distances between objects in various clusters. Additionally, some index employ metrics based on density.

The following is the basic method to use validation methods to identify the best partition and equally useful number of a set of items.

Step 1: Create a list of clustering methods that will be used on the data set.

Step 2: Use several parameter combinations for each clustering technique to obtain various clustering outcomes.

Step 3: For each division you received in Step 2, calculate the matching validation index.

Step 4: Determine the ideal cluster size and appropriate partition based on the criteria

Commonly used validation measures are shown in Table I. To the best of our knowledge, these measures cover a substantial portion of the validation measures that are accessible in several disciplines, including machine learning, data mining, and information retrieval. The computation forms for the measurements are provided in the "definition" column. Then, we simply describe these metrics. Including *DB, XB*, and *SIL*, take into account both evaluation criteria (compactness and separation) in the form of a ratio. However, certain index, such *RMSSTD, RS*, and *Γ*, only take into account one factor. The square root of the pooled sample variance for all the attributes determines the Root-mean-Square Standard Deviation (RMSSTD) [6]. It gauges how uniform the clusters that have developed. The ratio of the sum of squares within clusters to the sum of squares throughout the entire data set is known as R-Squared (RS). It gauges how different one cluster is from the others [6, 7]. By measuring the discrepancies between pairs of data items in two partitions, the Modified Hubert statistic Γ [8] assesses the difference between clusters. The average between- and within-cluster sum of squares is used as the basis for the Calinski-Harabasz index (CH) [9] evaluation of cluster validity.

Index I (I) [1] assesses compactness based on the sum of distances between items and their cluster centre and measures separation based on the maximum distance between cluster centres. The Dunn's index (D) [10] measures the inter-cluster separation as the smallest pairwise distance between objects in distinct clusters and the intracluster compactness as the largest diameter among all clusters. These three index have the formula: index = (a.separation)/ (b.compactness), where b and a are weights. Maximizing the value of these index yields the ideal cluster number.

The pairwise difference of between-cluster and within-cluster distances is used by the Silhouette Index (SIL) [11] to validate the clustering performance. Additionally, minimising the value of this index yields the ideal cluster number. Rousseeuw (1987) suggested the silhouette statistic for assessing clusters and figuring out the ideal number. Let a (i) represent the average dissimilarity between objects in the i[th] object's cluster, and b (i) represent the average dissimilarity between objects in the nearby cluster, which is defined as the cluster with the lowest average dissimilarity.

The Davies-Bouldin index (DB) [12] gets computed. The highest number is given to each cluster *C* as its cluster similarity, and the similarities between each cluster and each other cluster are calculated for each cluster. Then, by averaging all of the cluster similarities, the DB index may be created. The clustering outcome is better the smaller the index. The best partition is made by minimising this index, which makes clusters the most distinct from one another.

According to the Xie-Beni index (XB) [13], the intra-cluster compactness is the mean square distance between each data object and its cluster centre, and the inter-cluster separation is the minimum square distance between cluster centres. When the minimum of *XB* is discovered, the ideal cluster number is attained. Kim *et al.* [14] proposed the index *DB* and *XB* as upgrades to *DB* and *XB.* We shall apply these two enhanced measures in this paper

The principles of average scattering and total cluster separation serve as the foundation for the SD index (SD) [15] notion. According to the variances of the cluster objects, the first term assesses compactness, and the second term assesses separation difference according to the separations between cluster centres. The total of these two variables determines the value of the index, and decreasing the value of *SD* will yield the ideal number of clusters. Density is taken into account by the Silhouette index (*SIL*) [16] to calculate inter-cluster separation. The fundamental principle is that at least one of the densities of each pair of cluster centres should be higher than the density of the other centre. Similar to *SD*, the intra-cluster compactness is present. The maximum value of *SIL* denotes the ideal cluster number, and the index is the sum of these two terms.

Other validation techniques are described in the literature [17–20]. Some, however, perform poorly, and others are made for data sets with particular arrangements. Consider the Symmetry distance-based index (Sym-index) and Composed Density (CD) between and within clusters index as examples. Finding the representatives for each cluster is challenging for *CD*, which leads to an unstable outcome. Additionally, only data sets with intrinsic symmetry can be handled by Sym-index. As a result, for the remainder of the study, we concentrate on the aforementioned 10 validation measures. And we'll abbreviate each of these metrics throughout this paper.

## III. Understanding of Clustering Validation Measures

In this section, we offer a thorough analysis of the 10 validation measures described in Section II and look into the various features of each measure's validation properties, which may be useful for choosing an index. If not stated, the experiment's clustering algorithm is K-means [21] implemented by CLUTO [22].

### *A. Effects of Monotonicity*

The following experiment can be used to assess the monotonicity of various validation index. We run the K-means method on the well-separated data set and obtain the clustering results for various cluster densities.

According to Fig. 1, well separated is a synthetic data set made up of five clusters that are well-separated. The first three index rise or decrease monotonically as the cluster number *NC* increases, according to the experiment's results (Table 2). The remaining seven index, on the other hand, attain their maximum or minimum value when *NC* equals the actual cluster number. The first three index' monotonicity can be explained by a few different factors. SSE (Sum of Square Error), which is a measure of error, reduces as NC raises. Since $NC << n$ in reality, *n-NC* can be thought of as a constant number. As a result, *RMSSTD* reduces as *NC* raises. Additionally, we have *RS = (TSS -SSE)/T SS (TSS - Total Sum of Squares)* and *TSS = SSE + SSB (SSB - Between group Sum of Squares)*, both of which are constants for a given collection of data. As a result, *RS* raises as *NC* raises.

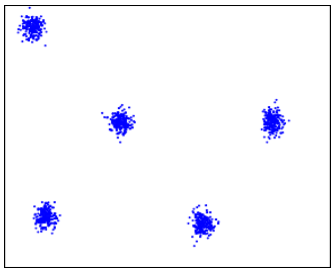


Fig. 1. The data set wellseparated.

Table 2. Results of monotonicity, true NC = 5

| | 2 | 3 | 4 | **5** | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|
| **RMS STD** | 28.49 | 20.80 | 14.82 | **3.201** | 3.081 | 2.957 | 2.834 | 2.715 |
| **RS** | 0.627 | 0.801 | 0.899 | **0.994** | 0.995 | 0.996 | 0.996 | 0.997 |
| **Γ** | 2973 | 3678 | 4007 | **4342** | 4343 | 4344 | 4346 | 4347 |
| **CH** | 1683 | 2016 | 2968 | **52863** | 45641 | 41291 | 38580 | 36788 |
| **I** | 3384 | 5759 | 11230 | **106163** | 82239 | 68894 | 58420 | 50259 |
| **D** | 0.491 | 0.549 | 0.58 | **2.234** | 0.025 | 0.017 | 0.009 | 0.01 |
| **SIL** | 0.607 | 0.707 | 0.004 | **0.825** | 0.718 | 0.579 | 0.475 | 0.391 |
| **DB** | 0.716 | 0.683 | 0.522 | **0.12** | 0.521 | 0.803 | 1.016 | 1.168 |
| **SD** | 0.215 | 0.124 | 0.075 | **0.045** | 0.504 | 0.486 | 0.538 | 0.553 |
| **XB** | 0.265 | 0.374 | 0.495 | **0.254** | 35.099 | 35.099 | 36.506 | 38.008 |

Only data objects in various clusters will be counted in the equation according to the definition of Γ. The number of items in each cluster will be n/2, and the number of distance pairs will actually be $n^2/4$ if the data set is partitioned into two equal clusters. When the data set is split into three equal clusters, $n^2/3$ pairs of distances will be counted for each cluster, which will contain n/3 objects. As a result, as the cluster number NC improves, more pairs of distances are measured, increasing the value of Γ. Further investigation reveals that these three index only consider separation or compactness, respectively. Only separation is taken into account by RS and *Γ*, while compactness is the only factor in *RMSSTD*. The *RMSSTD*, *RS*, and *Γ* curves will all exhibit monotonicity, which causes them to be either upward or

downward sloping. The elbow, or shift point of the curves, is said to be the location where the ideal cluster number is reached [7]. We won't go into these three index in the following sections, though, because determining the shift point is difficult and very subjective.

### B. Effects of Noise

We have the following experiment using the well separated noise data set to assess the impact of noise on validation index. As seen in Fig. 2, the synthetic data collection well separated. Noise was created by adding 5% noise to the original data set Well separated. Table 3 displays the cluster numbers chosen using the index.

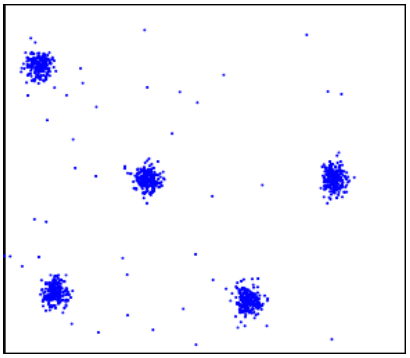


Fig. 2. The data set well separated-noise.

Table 3. Results of noise, true NC =5

| | 2 | 3 | 4 | **5** | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|
| **CH** | 1626 | 1846 | 2554 | 10174 | **14677** | 12429 | 11593 | 11088 |
| **I** | 3213 | 5073 | 9005 | **51530** | 48682 | 37568 | 29693 | 25191 |
| **D** | 0.0493 | 0.0574 | **0.0844** | 0.0532 | 0.0774 | 0.0682 | 0.0692 | 0.0788 |
| **SIL** | 0.59 | 0.67 | 0.783 | **0.802** | 0.025 | 0.653 | 0.626 | 0.596 |
| **DB** | 0.739 | 0.721 | 0.56 | **0.18** | 0.508 | 0.71 | 0.863 | 0.993 |
| **SD** | 0.069 | 0.061 | 0.05 | **0.045** | 0.046 | 0.055 | 0.109 | 0.121 |
| **XB** | 0.264 | 0.38 | 0.444 | **0.251** | 0.445 | 0.647 | 2.404 | 3.706 |

The outcomes of the experiment demonstrate that *D* and *CH* picked the incorrect cluster number. According to analysis, there are a few reasons why noise has a big impact on *D* and *CH*.

D measures the inter-cluster separation as the smallest pairwise distance between items in distinct clusters $\left(min_{x\in C_i, y\in C_j}\, d(x,y)\right)$ and the intra-cluster compactness as the largest diameter across all clusters $\left(max_k \left\{max_{x,y\in C_k} d(x,y)\right\}\right)$ and maximizing D's value will yield the ideal number of clusters. Since it only employs the minimum pairwise distance, rather than the average pairwise distance, across objects in different clusters, the inter-cluster separation can decrease abruptly when noise is present. As a result, the noise may have an impact on the value of D and the accompanying ideal cluster number. Since *CH = (SSB/SSE). ((n-NC)/ (NC-1)) and ((n-NC)/ (NC-1))* are constants for the same *NC*, we can only concentrate on the (SSB/SSE) portion. Compared to SSB, SSE grows more noticeably when noise is present. As a result, for a given *NC*, *CH* will drop due to the influence of noise, making CH's value unstable. Finally, noise will have an impact on the ideal cluster size. In addition, the other index will be less sensitively affected by noise than CH and D. We can see that the values of other index more or less vary when comparing Table 2 to Table 3. The ideal cluster number proposed by *I* will likewise be inaccurate if we add 20% noise to the well-separated data set. Therefore, in practice, it is generally advisable to reduce noise before clustering in order to limit the negative effects of noise.

### C. Effects of Density

Several clustering algorithms find it difficult to work with data sets with varying densities. We are therefore quite curious to know if it also influences the results of the validation measures. On a created data set called Different density, which has varied density, an experiment is run. Only *I* proposes the incorrect ideal cluster number, according to the findings presented in Table 4. Fig. 3 depicts the Different density in detail.

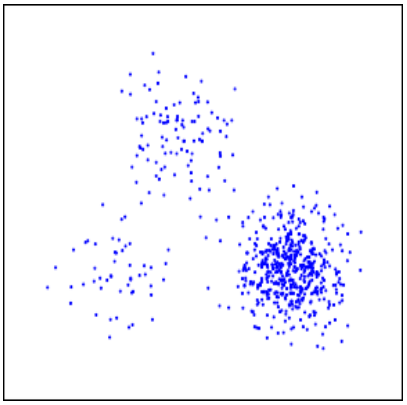


Fig. 3. The data set different density.

Table 4. Results of density, true NC = 3

| | 2 | **3** | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|
| ***CH*** | 1172 | **1197** | 1122 | 932 | 811 | 734 | 657 | 591 |
| ***I*** | **120.1** | 104.3 | 93.5 | 78.6 | 59.9 | 56.1 | 44.8 | 45.5 |
| ***D*** | 0.0493 | **0.0764** | 0.0048 | 0.0049 | 0.0049 | 0.0026 | 0.0026 | 0.0026 |
| ***SIL*** | 0.372 | **0.587** | 0.463 | 0.275 | 0.312 | 0.278 | 0.244 | 0.236 |
| ***DB*** | 0.658 | **0.498** | 1.001 | 1.186 | 1.457 | 1.688 | 1.654 | 1.696 |
| ***SD*** | 0.705 | **0.371** | 0.672 | 0.692 | 0.952 | 1.192 | 1.103 | 1.142 |

It is difficult to determine why ***I*** did not provide the correct cluster number. We can see that ***I*** keep falling as NC for the clusters rises. According to our hypothesis, one reason could be the K-means algorithm's uniform effect, which tends to

divide items into groups of roughly equal sizes [23].

***I*** calculate compactness by adding the distances between each object and the centre of the cluster. When *NC* is small, it is likely that items with a high density are in the same cluster, causing the total of distances to almost remain constant. The overall sum won't vary too much because the majority of the objects are in a single cluster. As a result, because NC is in the denominator, ***I*** will drop as *NC* rises.

### *D. Effect of Subclusters*

Clusters that are near one another are referred to as subclusters. Four of the five clusters in the synthetic data set Subclusters shown in Fig. 4 are subclusters because they can be combined to generate two pairs of clusters. The experiment's findings, which are shown in Table 5, assess the validation measures ability to handle data sets with subclusters. *D*, *SIL*, *DB*, *SD*, and *XB* provide the incorrect ideal cluster numbers for the Subclusters data set, while *I*, *CH*, and *SIL* offer the right ones.

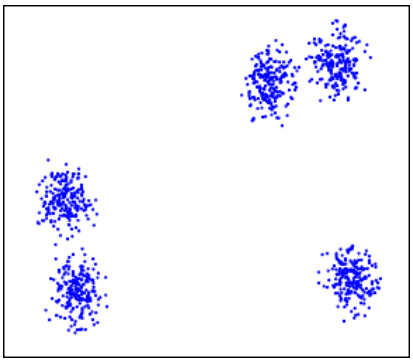


Fig. 4: The data set subcluster.

Table 5. Results of subclusters, true NC=5

| | 2 | 3 | 4 | **5** | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|
| ***CH*** | 3474 | 7851 | 8670 | **16630** | 14310 | 12900 | 11948 | 11354 |
| ***I*** | 2616 | 5008 | 5594 | **9242** | 7021 | 5745 | 4803 | 4248 |
| ***D*** | 0.741 | **0.7864** | 0.0818 | 0.0243 | 0.0243 | 0.0167 | 0.0167 | 0.0107 |
| ***SIL*** | 0.736 | 0.709 | 0.026 | **0.737** | 0.587 | 0.49 | 0.402 | 0.325 |
| ***DB*** | 0.445 | **0.353** | 0.54 | 0.414 | 0.723 | 0.953 | 1.159 | 1.301 |
| ***SD*** | 0.156 | **0.096** | 0.164 | 0.165 | 0.522 | 0.526 | 0.535 | 0.545 |
| ***XB*** | 0.378 | **0.264** | 1.42 | 1.215 | 12.538 | 12.978 | 14.037 | 14.858 |
| ***XB*** | 0.408 | **0.313** | 3.188 | 3.078 | 6.192 | 9.082 | 8.897 | 8.897 |

When the cluster number changes from $NC_{optimal}$ to $NC_{optimal+1,}$ intercluster separation should naturally reduce [14]. However, at $NC < NC_{optimal}$, declines can be seen for *D*, *DB*, *SD*, and *XB*. The causes are listed below. SIL determines the inter-cluster separation by averaging the smallest distances between clusters. When nearby subclusters are treated as a single large cluster in a data set with subclusters, the inter-cluster separation will be at its greatest value. Due to subclusters, the incorrect optimal cluster number will be selected. As a measure of separation, *XB* employs the smallest pairwise distance between cluster centres. When adjacent subclusters are treated as one huge cluster in a data set with subclusters, the measure of separation will reach its maximum value. As a result, utilizing *XB* won't help you find the right cluster number. Due to space constraints, we won't go into detail about the reasons for *D*, *SD*, and *DB* here, but they are fairly similar to the explanation for *XB*.

### *E. Effect of Skewed Distributions*

In a data set, it is typical for clusters to be of different sizes. A fictitious data set Skewdistribution with skewed distributions is shown in Fig. 5. There is one substantial cluster and two smaller ones. K-means performs poorly when dealing with skewed distributed data sets because it has the uniform effect, which tends to partition objects into nearly equal sizes [23].

We use four popular algorithms from four distinct categories to illustrate this claim, including K-means (prototype-based), DBSCAN (density-based), Agglo based on average link (hierarchical) [2], and Chameleon (graph-based) [25]. Since three is the actual cluster number, we apply each of them to the Skewdistribution and divide the data set into three clusters. Fig. 6 demonstrates that Chameleon performs the best while K-means performs the poorest.

To assess how well various index perform on data sets with skewed distributions, an experiment is conducted on the data set Skew distribution. The clustering algorithm that we employ is Chameleon. The experiment's findings, which are presented in Table 6, demonstrate that only *CH* is unable to provide the ideal cluster number. Given that *CH = (T SS/SSE -1). ((n -NC)/ (NC -1))* and *TSS* is a constant number of a particular data collection Therefore, *CH* is fundamentally founded on SSE, which also serves as the foundation for the K-means algorithm. K-means cannot handle skewed distributed data sets, as was already mentioned. The same conclusion therefore holds true for *CH.*

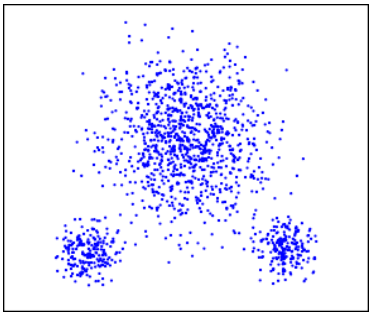
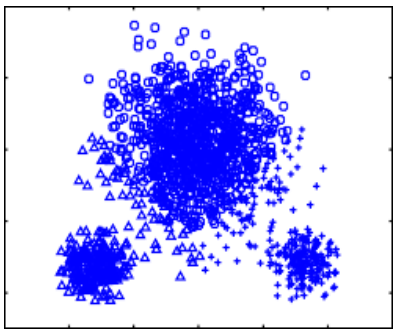
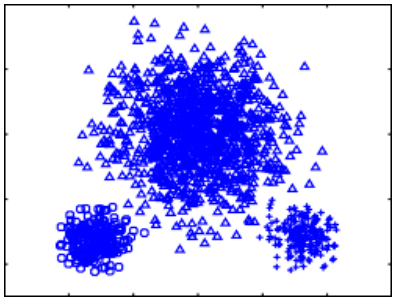


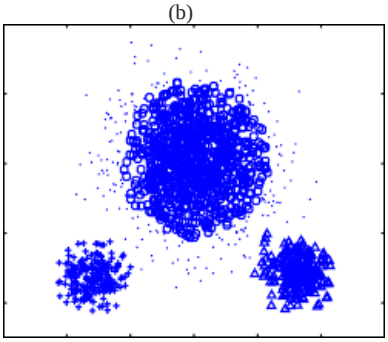
Fig. 5. The data set skew distribution.



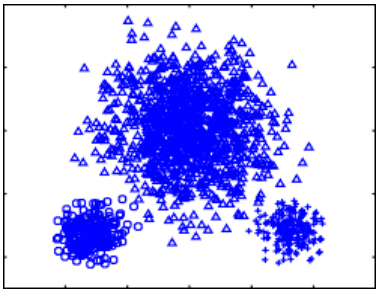
(a)



(b)



(c)



(d)

Fig. 6. (a) K-means (b) Agglomerative (c) DBSCAN, (d) Chameleon; Clustering results on the data set Skew distribution by different algorithms where NC = 3

The validation qualities of each of the 10 validation measures in each of the five elements examined above are listed in Table 7, which may provide direction for actual index selection. In this table, "_" specifies a property that has not been tested and "*" indicates a scenario that cannot be handled. As seen in Table 7, SIL is the only validation measure that performs well in each of the five criteria, although the other measures have some limits in various contexts, primarily in relation to noise and subclusters.

Table 6. Results of skewed distributions true NC = 3

| | 2 | **3** | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|
| ***CH*** | 788 | 1590 | 1714 | **1905** | 1886 | 1680 | 1745 | 1317 |
| ***I*** | 232.3 | **417.9** | 334.5 | 282.9 | 226.7 | 187.1 | 172.9 | 125.5 |
| ***D*** | 0.0286 | **0.0342** | 0.0055 | 0.0069 | 0.0075 | 0.0071 | 0.0075 | 0.0061 |
| ***SIL*** | 0.486 | **0.621** | 0.301 | 0.538 | 0.457 | 0.371 | 0.37 | 0.309 |
| ***DB*** | 0.571 | **0.466** | 0.844 | 0.807 | 0.851 | 1.181 | 1.212 | 1.875 |
| ***SD*** | 0.327 | **0.187** | 0.294 | 0.274 | 0.308 | 0.478 | 0.474 | 0.681 |
| ***XB*** | 0.369 | **0.264** | 1.102 | 0.865 | 1.305 | 3.249 | 3.463 | 7.716 |

Table 7. Performance of different index

| | *RMSSTD* | *RS* | Γ | *CH* | *I* | *D* | *SIL* | *DB* | *SD* | *XB* |
|---|---|---|---|---|---|---|---|---|---|---|
| **Mono.** | * | * | * | | | | | | | |
| **Noise** | _ | _ | _ | * | | * | | | | |
| **Dens.** | – | – | – | | * | | | | | |
| **Subc.** | – | – | – | | | * | | * | * | * |
| **Skew Dis.** | – | – | – | * | | | | | | |

## IV. Conclusion

In this study, we looked into five different elements of a set of 10 internal clustering validation metrics for crisp clustering: monotonicity, noise, density, subclusters, and skewed distributions. The 10 validation metrics were assessed using computational experiments on five synthetic data sets, each of which accurately represents one of the five elements mentioned above. The experiment's findings show that the majority of the current measurements have certain limitations in various application settings. *SIL* is the only indicator that excels in each of the five criteria. The validation qualities of these 10 validation measures are summarized in Table 7, which may be used as assistance when choosing an index in actual use.

## Conflict of Interest

The authors declare no conflict of interest.

## Author Contributions

Urvashi Soni Conducted the research, Sunita Dwivedi wrote the paper; all authors had approved the final version.

## References

[1] U. Maulik and S. Bandyopadhyay, "Performance evaluation of some clustering algorithms and validity index," *IEEE PAMI*, vol. 24, pp. 1650–1654,

[2] A. K. Jain and R. C. Dubes, "Algorithms for clustering data," *Upper Saddle River*, NJ, USA: Prentice-Hall, Inc.,1988.

[3] J. Wu, H. Xiong and J. Chen, "Adapting the right measures for k-means clustering," *KDD*, pp. 877–886, 2009.

[4] P. N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*, USA: Addison-Wesley Longman, Inc., 2014.

[4] Y. Zhao and G. Karypis, "Evaluation of hierarchical clustering algorithms for document datasets," in *Proc. of CIKM*, pp. 515–524, 2022.
[5] S. Sharma, *Applied Multivariate Techniques*, New York, NY, USA: John Wiley & Sons, Inc., vol. 2, 2008.
[6] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On clustering validation techniques," *Journal of Intelligent Information Systems*, vol. 17, pp. 107–145, 2001.
[7] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, no. 1, pp. 193–218, 1985.
[8] T. Calinski and J. Harabasz, "A dendrite method for cluster analysis," *Comm. in Statistics*, vol. 3, no. 1, pp. 1–27, 1974.
[9] J. Dunn, "Well separated clusters and optimal fuzzy partitions," *J. Cybern*., vol. 4, no. 1, pp. 95–104, 1974.
[10] P. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math*., vol. 20, no. 1, pp. 53–65, 1987.
[11] D. Davies and D. Bouldin, "A cluster separation measure," *IEEE PAMI*, vol. 1, no. 2, pp. 224–227, 1979.
[12] X. L. Xie and G. Beni, "A validity measure for fuzzy clustering," IEEE PAMI, vol. 13, no. 8, pp. 841–847, 1991.
[13] M. Kim and R. S. Ramakrishna, "New index for cluster validity assessment," *Pattern Recogn. Lett.*, vol. 26, no. 15, pp. 2353–2363, 2005.
[14] M. Halkidi, M. Vazirgiannis, and Y. Batistakis, "Quality scheme assessment in the clustering process," in *Proc. PKDD*, London, UK, pp. 265–276, 2000.
[15] M. Halkidi and M. Vazirgiannis, "Clustering validity assessment: Finding the optimal partitioning of a data set," in Proc. ICDM, Washington, DC, USA, pp. 187–194, 2001.
[16] S. Saha and S. Bandyopadhyay, "Application of a new symmetry-based cluster validity index for satellite image segmentation," *IEEE Geoscience and Remote Sensing Letters*, 2008.
[17] M. Halkidi and M. Vazirgiannis, "Clustering validity assessment using multi-representatives," in *Proc. Hellenic Conference on Artificial Intelligence*, 2002.
[18] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *J. Royal Statistical Society*, vol. 63, no. 2, pp. 411–423, 2001.
[19] B. S. Y. Lam and H. Yan, "A new cluster validity index for data with merged clusters and different densities," in *Proc. IEEE ICSMC*, 2005, pp. 798–803.
[20] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. BSMSP*, 1967, pp. 281–297.
[21] G. Karypis, Cluto – Software for clustering high-dimentional datasets," *Version*, vol. 2, no. 2, 2006.
[22] H. Xiong, J. Wu, and J. Chen, "K-means clustering versus validation measures: a data distribution perspective," in *Proc.* KDD, New York, NY, USA, 2006, pp. 779–784.
[23] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, "A densitybased algorithm for discovering clusters in large spatial databases with noise," in *Proc.* KDD, 1996, pp. 226–231.
[24] G. Karypis, E.-H. S. Han, and V. Kumar, "Chameleon: Hierarchical clustering using dynamic modeling," *Computer*, vol. 32, no. 8, pp. 68–75, 1999.