A New Index for Clustering Evaluation Based on Density Estimation

Gangli Liu
Tsinghua University
gl-liu13@mails.tsinghua.edu.cn

ABSTRACT

A new index for internal evaluation of clustering is introduced. The index is defined as a mixture of two sub-indices. The first sub-index $I_a$ is called the Ambiguous Index; the second sub-index $I_s$ is called the Similarity Index. Calculation of the two sub-indices is based on density estimation to each cluster of a partition of the data. An experiment is conducted to test the performance of the new index, and compared with six other internal clustering evaluation indices – Calinski-Harabasz index, Silhouette coefficient, Davies-Bouldin index, CDbw, DBCV, and VIASCKDE, on a set of 145 datasets. The result shows the new index significantly improves other internal clustering evaluation indices.

KEYWORDS

Clustering Evaluation; Calinski-Harabasz Index; Silhouette Coefficient; Davies-Bouldin Index; Kernel Density Estimation

1 INTRODUCTION

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters). Due to its intensive applications in data analysis, people have developed a wide variety of clustering algorithms. Such as K-Means, Gaussian mixtures, Spectral clustering [16], Hierarchical clustering [11], DBSCAN [20], BIRCH [28] etc. With so many choices of clustering algorithms, it is vital to have good tools to recognize it when a clustering algorithm has done a good job.

Evaluation (or “validation”) of clustering results is as difficult as the clustering itself [19]. Popular approaches involve “internal” evaluation and “external” evaluation. In internal evaluation, a clustering result is evaluated based on the data that was clustered itself. Popular internal evaluation indices are Davies-Bouldin index [18], Silhouette coefficient [1], Dunn index [3], and Calinski-Harabasz index [14] etc. In external evaluation, the clustering result is compared to an existing “ground truth” classification, such as the Rand index [26]. However, “knowledge of the ground truth classes is almost never available in practice” [17]. In this article, we introduce a new index that belongs to the “internal evaluation” category of clustering evaluation, it is based on density estimation.

2 RELATED WORK

Calinski-Harabasz index [14], Silhouette coefficient [1], and Davies-Bouldin index [18] are three of the most popular techniques for internal clustering evaluation.

2.1 Calinski-Harabasz index (CH)

For a set of data $E$ of size $n_E$ which has been clustered into $k$ clusters, the Calinski-Harabasz score $s$ is defined as the ratio of the between-clusters dispersion mean and the within-cluster dispersion: [17]

$$s = \frac{\text{tr}(B_k)}{\text{tr}(W_k)} \times \frac{n_E - k}{k - 1}$$

where $\text{tr}(B_k)$ is trace of the between group dispersion matrix and $\text{tr}(W_k)$ is the trace of the within-cluster dispersion matrix defined by:

$$W_k = \frac{1}{k} \sum_{q=1}^{k} \sum_{x \in C_q} (x - c_q)(x - c_q)^T$$

$$B_k = \frac{1}{k} \sum_{q=1}^{k} n_q(c_q - c_E)(c_q - c_E)^T$$

with $C_q$ the set of points in cluster $q$, $c_q$ the center of cluster $q$, $c_E$ the center of $E$, and $n_q$ the number of points in cluster $q$.

A higher Calinski-Harabasz score relates to a model with better defined clusters (Table 1).

2.2 Silhouette coefficient (SC)

The Silhouette coefficient for a single sample is given as:

$$s = \frac{b - a}{\max(a, b)}$$

where $a$ is the mean distance between a sample and all other points in the same class. $b$ is the mean distance between a sample and all other points in the next nearest cluster. The Silhouette coefficient for a set of samples is given as the mean of Silhouette coefficient for each sample. The range of Silhouette coefficient is $[-1, 1]$. A higher Silhouette coefficient score relates to a model with better defined clusters.

2.3 Davies-Bouldin index (DB)

The Davies-Bouldin index can be calculated by the following formula:

$$DB = \frac{1}{n} \sum_{i=1}^{n} \max_{j \neq i} \left( \frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right)$$

1https://en.wikipedia.org/wiki/Cluster_analysis
where \( n \) is the number of clusters, \( c_i \) is the centroid of cluster \( i \), \( \sigma_i \) is the average distance of all elements in cluster \( i \) to centroid \( c_i \), and \( d(c_i, c_j) \) is the distance between centroids \( c_i \) and \( c_j \). The clustering algorithm that produces a collection of clusters with the smallest Davies-Bouldin index is considered the best algorithm based on this criterion.

### 2.4 Other internal evaluation indices

Besides the three popular internal clustering evaluation indices, numerous criteria have been proposed in the literature [2, 4, 15, 23, 27]. Such as DBCV [15], Xie-Beni (XB) [24], CDbw [8], S_Dbw [7], and RMSSTD [6]. Besides, new cluster validity indices keep emerging, such as the CVNN [13], CVDD [10], DSI [5], SCV [25], AWCD [12] and VIASCCKDE [21].

It is impractical to compare with all the indices listed above. Therefore, we select six of above internal clustering evaluation indices – Calinski-Harabasz index, Silhouette coefficient, Davies-Bouldin index, CDbw, DBCV, and VIASCCKDE, to compare with the new index (see Section 4). Among them, Calinski-Harabasz index, Silhouette coefficient, and Davies-Bouldin index are well known popular internal indices. CDbw and DBCV are two popular density-based internal clustering evaluation indices. VIASCCKDE is a newly published index.

### 2.5 Kernel density estimation

In probability and statistics, density estimation is the construction of an estimate, based on observed data, of an unobservable underlying probability density function [22].

Kernel density estimation (KDE) is a non-parametric way to estimate the probability density function of a random variable. Kernel density estimation is a fundamental data smoothing problem where inferences about the population are made, based on a finite data sample.  

Let \( (x_1, x_2, ..., x_n) \) be independent and identically distributed samples drawn from some univariate distribution with an unknown density \( f \) at any given point \( x \). We are interested in estimating the shape of this function \( f \). Its kernel density estimator is:

\[
\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)
\]

where \( K \) is the kernel – a non-negative function – and \( h > 0 \) is a smoothing parameter called the bandwidth. A kernel with subscript \( h \) is called the scaled kernel and defined as \( K_h(x) = 1/hK(x/h) \). A range of kernel functions are commonly used: uniform, triangular, biweight, triweight, Epanechnikov, normal, and others.

### 3 Calculation of the New Index

#### Definition 3.1. The new index

Suppose dataset \( \Omega \) of size \( n_\Omega \) has been clustered into \( K \) clusters, the new index \( I \) is defined as a mixture of two sub-indices:

\[
I = \delta I_s + (1 - \delta) I_a
\]

where \( \delta \) is a hyper-parameter in range \([0, 1]\) which can be optimized by a training dataset. The first sub-index \( I_s \) is called the Ambiguous Index; the second sub-index \( I_a \) is called the Similarity Index. The range of the new index \( I \) is \([0, 1]\), and it is “the smaller, the better”.

To calculate the two sub-indices, we first fit a density estimator \( \hat{d}_q \) for each cluster \( C_q \).

#### 3.1 The Ambiguous Index

To calculate the Ambiguous Index \( I_a \), we first determine the Territory of Cluster \( q \).

**Definition 3.2. Territory of Cluster \( q \)**

The Territory of Cluster \( q \), noted \( T_q \), is calculated with following equations:

\[
G_q = \{ \log(d_q(x)) \mid x \in C_q \}
\]

\[
T_q = [\min(G_q) - \alpha_1 \Delta_G, \max(G_q) + \alpha_2 \Delta_G]
\]

where \( d_q(x) \) is the likelihood of \( x \) estimated by density estimator \( \hat{d}_q \). \( \log(d_q(x)) \) is its log-likelihood. \( \Delta_G \) is Standard Deviation (SD) of \( G_q \). \( \alpha_1 \) and \( \alpha_2 \) are two constant hyper-parameters. If \( \Delta_G \) equals zero, then replace \( \alpha_1 \times \Delta_G \) and \( \alpha_2 \times \Delta_G \) with two constants \( \beta_1 \) and \( \beta_2 \). The hyper-parameters can be optimized by a training dataset.

\( T_q \) is a range for log-likelihood. To determine if a data point \( y \) is in the Territory of Cluster \( q \), we calculate the log-likelihood of \( y \) with the density estimator \( \hat{d}_q \). If \( \log(d_q(y)) \) is in the range of \( T_q \), then \( y \) is in the Territory of Cluster \( q \).

**Definition 3.3. Ambiguous Point**

If a data point is in the territories of at least two clusters, it is defined as an Ambiguous Point.

**Definition 3.4. Ambiguous Index**

The Ambiguous Index is defined as the ratio of number of Ambiguous Points in \( \Omega \) to total number of points in \( \Omega \).

\[
I_a = \frac{a_\Omega}{n_\Omega}
\]

The range of Ambiguous Index \( I_a \) is \([0, 1]\). Figure 1, 2, and 3 illustrate an example of calculating the Ambiguous Index. The dataset is from a previous paper [9]. A clustering algorithm partitions the dataset into four clusters. The points tagged with a red cross in Figure 3 are Ambiguous Points, according to Definition 3.3.
Table 1: Higher is better or smaller is better?

| CH | SC | DB | CDbw | DBCV | VIASCKDE | The new index |
|----|----|----|------|------|-----------|---------------|
| higher | higher | smaller | higher | higher | higher | higher | smaller |

Table 2: URL of implementations of three density-based indices

| URL            | Description          |
|----------------|----------------------|
| CDbw           | https://pypi.org/project/cdbw/ |
| DBCV           | https://github.com/scikit-learn-contrib/hdbscan/blob/master/hdbscan/validity.py |
| VIASCKDE       | https://github.com/senolali/VIASCKDE |

3.2 The Similarity Index

The Similarity Index is calculated with the following equations:

**Definition 3.5. Similarity Index**

\[ L_q = \{ d_q(x) \mid x \in C_q \} \]  \hspace{1cm} (5)

\[ S_q = \frac{1}{\max(L_q)} \sum_{y \in L_q} y \]  \hspace{1cm} (6)

where \( d_q(x) \) is the likelihood of data point \( x \) evaluated by density estimator \( d_q \), \( L_q \) is the set of all points’ likelihood in cluster \( C_q \), \( \max(L_q) \) is the maximum value in \( L_q \). \( S_q \) is sum of likelihood normalized by the maximum likelihood in cluster \( C_q \). \( S_\Omega \) is sum of all normalized likelihood in data \( \Omega \). \( n_\Omega \) is the number of points in data \( \Omega \). \( \frac{S_\Omega}{n_\Omega} \) is the average normalized likelihood in \( \Omega \). Note the difference between likelihood and log-likelihood. Here we use likelihood.

The range of Similarity Index \( I_s \) is also \([0, 1]\).
We conducted an experiment to test the performance of the new algorithm with ward, complete, average, and single linkage (each linkage is evaluated by the scikit-learn project). Implementations of CDbw, DBCV, and Davies-Bouldin index are calculated with library functions provided by the scikit-learn project [17]. All the datasets are two-dimensional or three-dimensional, so that it is convenient to validate the outcomes by eyeballing it.

### 4.2 Settings of the experiment

For each dataset, seven clustering algorithms are used to partition the dataset with parameter K (number of clusters) ranging from 2 to 30. The seven clustering algorithms are: Agglomerative Clustering with ward, complete, average, and single linkage (each linkage is counted as a different clustering algorithm), Spectral Clustering, K-means Clustering, and Gaussian Mixture Clustering.

If two clustering algorithms’ partitioning of a dataset are the same, only one partitioning is kept. In addition, a reference partition (it is supposed to be the right partition) is provided for each dataset (see the first row of Figure 5).

Therefore, if there are no duplicate partitions and the reference partition is not in the seven clustering algorithms’ outcomes, then there are 204 (29 * 7 + 1 = 204) candidate partitions of a dataset to be evaluated by an index.

We used about 1/4 of the datasets (that is 36 of the 145 datasets) to select optimized values for the hyper-parameters in Equation 1 and 3. For density estimation of a cluster, we used the Kernel Density Estimation (KDE) library functions provided by the scikit-learn project. With the “bandwidth” hyper-parameter selected by Cross Validation and Grid Search, and the “kernel” fixed to be “gaussian”.

Then, we used the new index to evaluate the 204 (may be less due to duplicates) candidate partitions of each dataset, selecting the top 5 partitions ranked by the index. Then compared the results with the results evaluated by the six indices mentioned in Section 2. Scores of Calinski-Harabasz index, Silhouette coefficient, and Davies-Bouldin index are calculated with library functions provided by the scikit-learn project. Implementations of CDbw, DBCV, and VIASCKDE see Table 2.

### 4.3 Results of the experiment

Figure 5 is result for one dataset. The first row of the figure plots the dataset, tagged with dataset ID and the true number of clusters K, and the reference partition of the dataset. The following seven rows plot the top 5 partitions evaluated by each index (ranked from left to right). Each tagged with number of clusters K and scoring of the corresponding index, and the adjusted Rand index (AR) of the partition with the reference partition. The adjusted Rand index is used to evaluate a partition’s agreement with the reference partition of the dataset. “CH” stands for Calinski-Harabasz index, “SC” stands for Silhouette coefficient, “DB” stands for Davies-Bouldin index, and “T” is the new index. “VIASCKDE” is simplified to “VIA”.

To get a smaller index while the sixth row, the Ambiguous Index is calculated with library functions provided by the scikit-learn project [17]. All the datasets are two-dimensional or three-dimensional, so that it is convenient to validate the outcomes by eyeballing it.

### 5 WHY IT WORKS?

We use an example to intuitively show why the new index works. In the example, we use the Gaussian Mixture Model (GMM) to partition a dataset into, from $K = 2$ to $K = 31$ clusters, as illustrated in Figure 6.

Figure 4 shows how the two sub-indices and their sums varying along the number of clusters K. As mentioned in Section 3, the range of the new index $I$ is $[0, 1]$, and it is “the smaller, the better”. To get a smaller index $I$, the Ambiguous Index $I_A$ tries to pull the number of clusters $K$ to a smaller number; however, the Similarity Index $I_S$ tries to pull the number of clusters $K$ to a larger number. At some point, the two sub-indices get an equilibrium. The hyper-parameter in Equation 1 is to adjust the two sub-indices so that they are on comparable level.
Another example is illustrated in Figure 7 and 8. Figure 8 also shows how the result is affected by hyper-parameter $\delta$ in Equation 1. It can be seen that the result is not very sensitive to hyper-parameter $\delta$. When $\delta$ locates in an appropriate range, the new index will succeed in recognizing the right partition of the data.

An interesting question is: what if we use only one sub-index? For some easy recognition tasks, one sub-index alone can accomplish the recognition of the right partition; for some harder tasks, we need the cooperation of the two sub-indices.

| 8 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CH | S | S | F | F | F | F | F | F | F | F | S | F | F | F | F | F | F | F | F | F | F | F | F |
| SC | S | S | F | F | F | F | F | F | S | F | F | F | F | F | F | F | F | F | F | F | F | F | F |
| DB | S | S | F | F | F | F | F | F | S | F | F | F | F | F | F | F | F | F | F | F | F | F | F |
| CDbw | S | F | F | F | F | F | F | F | F | S | F | F | F | F | F | F | F | F | F | F | F | F | F |
| DBCV | F | F | S | F | S | F | S | F | F | F | F | S | F | F | F | F | F | F | F | F | F | F | F |
| VIA | F | F | F | F | S | F | F | F | F | F | F | F | F | F | F | F | F | F | F | F | F | F | F |
| New | S | S | S | S | S | S | S | S | S | S | S | S | S | S | S | S | S | S | S | S | S | S | S |
| 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 | 41 | 42 | 43 | 44 | 45 | 46 | 47 | 48 | 49 |
| CH | F | F | F | F | S | S | S | S | F | F | F | F | F | F | F | F | F | F | F | F | F | F | F |
| SC | F | F | F | F | F | F | F | F | F | F | F | F | F | F | F | F | F | F | F | F | F | F | F |
| DB | S | S | S | S | S | S | S | S | S | S | S | S | S | S | S | S | S | S | S | S | S | S | S |
| CDbw | F | F | F | F | S | S | S | S | F | F | F | F | F | F | F | F | F | F | F | F | F | F | F |
| DBCV | F | F | S | S | S | S | S | S | S | S | F | F | F | F | F | F | F | F | F | F | F | F | F |
| VIA | F | F | F | F | F | F | F | F | F | F | F | F | F | F | F | F | F | F | F | F | F | F | F |
| New | S | S | S | S | S | S | S | S | S | S | S | S | S | S | S | S | S | S | S | S | S | S | S |

Table 4: The seven indices’ performance on the 145 datasets. ‘S’: Succeeded, ‘F’: Failed, ‘New’: the new index

| CH | SC | DB | CDbw | DBCV | VIA | SCKDE | New |
|---|---|---|---|---|---|---|---|
| Accuracy | 27/145 | 38/145 | 42/145 | 8/145 | 56/145 | 11/145 | 74/145 |

Table 5: Accuracy of the seven indices

6 DISCUSSION

6.1 Several variants of the two sub-indices

There are several variants of the two sub-indices, which get comparable performance scores to the main sub-indices defined in Section 3, when tested on the datasets mentioned in Section 4.1.

6.1.1 Variant i of the Ambiguous Index, $I_{a_1i_1}$

To calculate $I_{a_1i_1}$, we first calculate the pairwise Ambiguous Index between each pair of clusters $i$ and $j$, noted $A_{ij}$, then compute the proportion of $A_{ij}$ that is positive.
6.1.2 Variant II of the Ambiguous Index, $I_{a,v2}$.
In $I_{a,v2}$, instead of computing the proportion of $A_{ij}$ that is positive, we compute the mean of $A_{ij}$ that is positive. If all $A_{ij}$ are zero, we just set the value to zero.

6.1.3 Variant III of the Ambiguous Index, $I_{a,v3}$.
To calculate $I_{a,v3}$, we first estimate the area of the territory for each cluster by using Monte Carlo method, then calculate the proportion of the area that has disputes to the total area of all clusters.

6.1.4 Variant I of the Similarity Index, $I_{s,v1}$.
To calculate $I_{s,v1}$, we first process $L_q$ in Equation 5 with a min-max normalization, then calculate the $S_q$ as the sum of the min-max normalized $I_q$.

6.1.5 Variant II of the Similarity Index, $I_{s,v2}$.
To calculate $I_{s,v2}$, we first calculate the center of $G_q$ in Equation 2. It can be the mean or median of $G_q$. Then calculate the distance (using absolute value) or squared-distance from each member of $G_q$ to the center. Then calculate mean of the distances. The log-likelihood can be replaced with likelihood. The distances can be min-max normalized.
Figure 6: 30 candidate partitions of a dataset

Figure 7: 30 candidate partitions of another dataset
6.2 The third sub-index, Boundary Index, $I_b$

To calculate Boundary Index $I_b$, we first determine the Boundary Points of cluster $C_q$.

**Definition 6.1. Boundary Point**

$$B_q = \min(G_q), \min(G_q) + \rho \cdot \Delta_G$$

where $G_q$ is defined in Equation 2, $\Delta_G$ is Standard Deviation (SD) of $G_q$, $\rho$ is a constant hyper-parameter. If a data point $p$'s log-likelihood evaluated by density estimator $d_q$, is in the range of $B_q$ then $p$ is defined as a Boundary Point of cluster $C_q$.

Figure 9 shows two clusters of data points with Boundary Points colored red.

**Definition 6.2. Boundary Index**

$$I_b = \frac{\sum_{i=1}^{K} N_{B_i}}{KN_\Omega}$$

where $N_{B_i}$ is the number of Boundary Points in cluster $i$, $N_\Omega$ is the total number of points in data $\Omega$, $K$ is the number of clusters.

Preliminary tests show a mixture of $I_a$, $I_s$, and $I_b$ can get a slightly better performance. However, the improvement is not significant. Besides, it increases model complexity.

6.4 Mixture of more ingredients

It is possible to use a mixture of more ingredient indices to get a better performance. E.g., some preliminary tests show a mixture of $I_a$, $I_s$, and $I_b$ can get a slightly better performance. However, the improvement is not significant. Besides, it increases model complexity.

6.5 Estimating the $\delta$ hyper-parameter

Maybe we should not fix the $\delta$ hyper-parameter in Equation 1 for all datasets, but to estimate it with characteristics of data $\Omega$. Such as number of points in $\Omega$, dimension of $\Omega$, global best bandwidth of $\Omega$ etc. However, how to calculate the $\delta$ need to be figured out.

6.6 Complexity

The complexity of the new index depends on what density estimation techniques are employed, to get a density estimator for each cluster $C_q$. It is generally slower than computation of non-density-based indices. Because we need to fit a density estimator for each cluster firstly. The computation can be accelerated by employing parallel computing.

6.7 Fixing random seeds

If randomness is involved in the process of fitting density estimators for each cluster, or selecting the best bandwidth for dataset $\Omega$, the value of the new index can be fixed by fixing the random seeds.

7 CONCLUSION AND FUTURE WORKS

We introduced a new index for internal evaluation of clustering, then tested and compared it with six popular or new internal clustering evaluation indices – Calinski-Harabasz index, Silhouette
REFERENCES

[1] S Aranganayagi and Kuttiyannan Thangavel. 2007. Clustering categorical data using silhouette coefficient as a relocating measure. In International conference on computational intelligence and multimedia applications (ICCAIMA 2007), Vol. 2. IEEE, 13–17.

[2] Ariel E Baya and Pablo M Granitto. 2013. How many clusters: A validation index for arbitrary-shaped clusters. IEEE/ACM Transactions on Computational Biology and Bioinformatics 10, 2 (2013), 401–414.

[3] James C Bezdek and Nikhil R Pal. 1995. Cluster validation with generalized Dunn’s indices. In Proceedings 1995 second New Zealand international two-stream conference on artificial neural networks and expert systems. IEEE Computer Society, 190–196.

[4] C-H Chou, M-C Su, and Eugene Lai. 2004. A new cluster validity measure and its application to image compression. Pattern Analysis and Applications 7, 2 (2004), 205–220.

[5] Shuyue Guan and Murray Loew. 2020. An internal cluster validity index using a distance-based separability measure. In 2020 IEEE 32nd international conference on tools with artificial intelligence (ICTAI). IEEE, 827–834.

[6] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. 2001. On clustering validation techniques. Journal of intelligent information systems 17, 2 (2001), 107–145.

[7] Maria Halkidi and Michalis Vazirgiannis. 2001. Clustering validity assessment: Finding the optimal partitioning of a data set. In Proceedings 2001 IEEE international conference on data mining. IEEE, 187–194.

[8] Maria Halkidi and Michalis Vazirgiannis. 2008. A density-based cluster validity approach using multi-representatives. Pattern Recognition Letters 29, 6 (2008), 773–786.

[9] Julia Handl and Joshua Knowles. 2004. Multiobjective clustering with automatic determination of the number of clusters. Technical Report (2004).

[10] Lianyu Hu and Caiming Zhong. 2019. An internal validity index based on density-involved distance. IEEE Access 7 (2019), 40038–40051.

[11] Stephen C Johnson. 1967. Hierarchical clustering schemes. Psychometrika 32, 3 (1967), 241–254.

[12] Qi Li, Shifeng Yue, Yaru Wang, Mingliang Ding, and Jia Li. 2020. A new cluster validity index based on the adjustment of within-cluster distance. IEEE Access 8 (2020), 202872–202885.

[13] Yanchi Liu, Zhongmou Li, Hui Xiong, Xuedong Gao, Junjie Wu, and Sen Wu. 2013. Understanding and enhancement of internal clustering validation measures. Pattern recognition 46, 1 (2013), 382–392.

[14] Ujjwal Maulik and Sagnamitra Bandyopadhyay. 2002. Performance evaluation of some clustering algorithms and validity indices. IEEE Transactions on pattern analysis and machine intelligence 24, 12 (2002), 1650–1654.

[15] Davoud Moulavi, Pablo A Jankowiak, Ricardo JCB Campello, Arthur Zimek, and Jörg Sander. 2014. Density-based clustering validation. In Proceedings of the 2014 SIAM international conference on data mining. SIAM, 839–847.

[16] Andrew Ng, Michael Jordan, and Yair Weiss. 2001. On spectral clustering: Analysis and an algorithm. Advances in neural information processing systems 14 (2001).

[17] F Pedregosa, G Varoquaux, A Gramfort, V Michel, B Thirion, O Grisel, M Blondel, P Prettenhofer, R Weiss, V Dubourg, J Vanderplas, A Passos, D Cournapeau, M Brucher, M Perrot, and E Duchesnay. 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12 (2011), 2825–2830.

[18] Slobodan Petrović. 2006. A comparison between the silhouette index and the davies-bouldin index in labelling ids clusters. (2006).

[19] Darius Piltzner, Richard Leibrandt, and David Powers. 2009. Characterization and evaluation of similarity measures for pairs of clusterings. Knowledge and Information Systems 19, 3 (2009), 361–394.

[20] Erich Schubert, Jörg Sander, Martin Ester, Hans Peter Kriegel, and Xiaowei Xu. 2017. DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. ACM Transactions on Database Systems (TODS) 42, 3 (2017), 1–21.

[21] Ali Şenol. 2022. VIACKDE Index: A Novel Internal Cluster Validity Index for Arbitrary-Shaped Clusters Based on the Kernel Density Estimation. Computational Intelligence and Neuroscience 2022 (2022).

[22] Bernard W Silverman. 2018. Density estimation for statistics and data analysis. Routledge.

[23] Lucas Vendramin, Ricardo JCB Campello, and Eduardo R Hruschka. 2010. Relative clustering validity criteria: A comparative overview. Statistical analysis and data mining: the ASA data science journal 3, 4 (2010), 209–235.

[24] Xuanli Lisa Xie and Gerardo Beni. 1991. A validity measure for fuzzy clustering. IEEE Transactions on pattern analysis and machine intelligence 13, 8 (1991), 841–847.

[25] Qin Xu, Qiang Zhang, Jimpie Li, and Bin Luo. 2020. Efficient synthetical clustering validity indexes for hierarchical clustering. Expert Systems with Applications 151 (2020), 113367.

[26] Ka Ye Yeung and Walter L Ruzzo. 2001. Details of the adjusted rand index and clustering algorithms, supplement to the paper an empirical study on principal component analysis for clustering gene expression data. Bioinformatics 17, 9 (2001), 763–774.

[27] Krista Rizman Žalik and Borut Žalik. 2011. Validity index for clusters of different sizes and densities. Pattern Recognition Letters 32, 2 (2011), 221–234.

[28] Tian Zhang, Raghu Ramakrishnan, and Miron Livny. 1996. BIRCH: An efficient data clustering method for very large databases. ACM sigmod record 25, 2 (1996), 103–114.

8 APPENDIX
```python
def kde_select_best_bandwidth_sliding_window(X):
    if len(X) == 1:
        best_bandwidth = {'bandwidth': 1.0}
        return best_bandwidth

    last_best_bandwidth = -1
    n = 0
    num_of_bandwidths = 10
    stepsize = 0.1

    while(True):
        logspace = np.logspace(n-stepsize, n, num_of_bandwidths)
        logspace = np.round(logspace, 10)
        global_best_bandwidth = kde_select_best_bandwidth(X, logspace)

        if np.abs(global_best_bandwidth['bandwidth'] - last_best_bandwidth) <= 1e-7:
            return global_best_bandwidth

        last_best_bandwidth = global_best_bandwidth['bandwidth']

    if global_best_bandwidth['bandwidth'] == np.min(logspace):
        n = np.round(n, 2)
        assert n > -10, "n > -10"
        elif global_best_bandwidth['bandwidth'] == np.max(logspace):
            n = np.round(n, 2)
            assert n < 10, "n < 10"
        else:
            return global_best_bandwidth

def kde_select_best_bandwidth(X, logspace):
    if len(X) < 5:
        cv = len(X)
    else:
        cv = 5
    cv_KFold = KFold(n_splits = cv, shuffle=True)
    params = {'bandwidth': logspace}
    grid = GridSearchCV(KernelDensity(), params, cv=cv_KFold, refit=False)
    grid.fit(X)
    best_bandwidth = grid.best_params_
    return best_bandwidth
```

Figure 10: Python code of sliding window technique
A New Index for Clustering Evaluation Based on Density Estimation

0000, 0000, Beijing, China
A New Index for Clustering Evaluation Based on Density Estimation

K = 28, CH = 413.648, AR = 0.273
K = 21, CH = 413.959, AR = 0.5291
K = 12, CH = 417.170, AR = 0.7013
K = 15, CH = 418.078, AR = 0.54641
K = 13, CH = 418.093, AR = 0.59040

K = 12, SC = 0.118, AR = 0.5712
K = 13, SC = 0.415, AR = 0.5336
K = 16, SC = 0.447, AR = 0.5661
K = 16, SC = 0.448, AR = 0.5305
K = 12, SC = 0.415, AR = 0.5435

K = 26, DC = 0.715, AR = 0.5365
K = 29, DC = 0.717, AR = 0.5257
K = 29, DC = 0.717, AR = 0.5346
K = 27, DC = 0.717, AR = 0.5265
K = 26, DC = 0.715, AR = 0.5209

K = 3, CDB = 0.107, AR = 0.9552
K = 4, CDB = 0.113, AR = 0.9558
K = 5, CDB = 0.114, AR = 0.9555
K = 5, CDB = 0.114, AR = 0.9555
K = 5, CDB = 0.114, AR = 0.9555

K = 2, D = 0.715, AR = 1.0
K = 29, DBC = 0.018, AR = 0.5289
K = 20, DBC = 0.457, AR = 0.5229
K = 30, DBC = 0.472, AR = 0.5215
K = 27, DBC = 0.472, AR = 0.5274

K = 4, VA = 0.117, AR = 0.6763
K = 3, VA = 0.127, AR = 0.7521
K = 3, VA = 0.104, AR = 0.7522
K = 3, VA = 0.104, AR = 0.43967
K = 2, VA = 0.104, AR = 2.8513

K = 25, TV = 0.315, AR = 0.5247
K = 26, TV = 0.315, AR = 0.5335
K = 23, TV = 0.315, AR = 0.5296
K = 25, TV = 0.315, AR = 0.5296
K = 21, TV = 0.315, AR = 0.5271
A New Index for Clustering Evaluation Based on Density Estimation
A New Index for Clustering Evaluation Based on Density Estimation
A New Index for Clustering Evaluation Based on Density Estimation

0000, 0000, Beijing, China
