A new cluster validity index for overlapping datasets

To cite this article: Erzhou Zhu et al 2019 J. Phys.: Conf. Ser. 1168 032070

View the article online for updates and enhancements.
A new cluster validity index for overlapping datasets

Erzhou Zhu1*, Xue Wang2 and Feng Liu3

1,2,3School of Computer Science and Technology, Anhui University, Hefei 230601, China

*Corresponding author’s e-mail: ezzhu@ahu.edu.cn

Abstract. The choice of the optimal clustering number \( K_{opt} \) is very important for forming the clustering results. As clustering is a kind of unsupervised learning method, it is still difficult to determine the \( K_{opt} \). The clustering validity index (CVI) is an effective method for determining the \( K_{opt} \) and evaluating the clustering results generated by clustering algorithms. However, many of the existing CVIs cannot efficiently with overlapping datasets. In this paper, we propose WCH, a new cluster validity index for overlapping datasets. The WCH index is constructed based on features of inner-cluster tightness, inter-cluster separation and inter-cluster overlapping of datasets. The WCH index can effectively process datasets with high overlapping and can accurately find the corresponding \( K_{opt} \). The comparison experiments are carried between our new WCH index and four existing CVIs on five simulation datasets. The experimental results have shown that the new WCH index is stable in forming the \( K_{opt} \) for many kinds of overlapping datasets.

1. Introduction

Clustering analysis belongs to the unsupervised learning method. It divides the samples of datasets into multiple clusters according to the similarity criteria[1]. Now, it is widely used in many fields such as data analysis, pattern recognition, and image processing[2]. Researches on clustering analysis are mainly focusing on two directions, clustering algorithms and cluster validity indexes.

Clustering algorithms are usually used to partition input datasets into clusters. Although clustering algorithms can effectively process datasets, it needs to find another way to determine the \( K_{opt} \) and evaluate the quality of clustering results. As an important method to evaluating the effect of clustering results generated by clustering algorithms, CVI has always been the focus of cluster analysis[3]. A CVI with good performance is a key to optimize and determine the \( K_{opt} \). At present, many CVIs have been proposed and been applied in different fields. Some of the commonly used CVIs are the Dunn's index (abbreviated by DI-index in this paper) [4], the Davies-Bouldin index (DBI-index) [5], the Bandyopadhy index (I-index) [6] and the Ibai Gurrutxaga index (COP-index)[7]. The DI-index is constructed based on the ratio of the smallest inter-cluster distance to the largest cluster diameter. The DBI-index is the ratio of the sum of the dispersions within the cluster to the sum of the separations between the clusters. The I-index measures the clustering result by mutual constraint of three factors. The COP-index is the ratio of the tightness within the cluster to the farthest adjacent distance. These CVIs can determine the optimal clustering numbers of datasets and evaluate assess the quality of clustering results. However, most of the existing CVIs are only applicable to non-overlapping datasets or deal datasets with low overlapping. For a dataset with high overlapping, many CVIs are stuck. This paper proposes WCH for effectively process overlapping datasets. The WCH index is constructed based on features of inner-cluster tightness, inter-cluster separation and inter-cluster overlapping of
datasets. The WCH index can effectively process datasets with high overlapping and can accurately find the corresponding $K_{opt}$.

2. The K-means algorithm

The K-means algorithm is simple to implement and widely used in many partitioning clustering algorithms. In this paper, the K-means algorithm is firstly used to cluster the dataset, and then the WCH is used to evaluate the clustering result and determine the $K_{opt}$. Actually, the K-means algorithm is a typical distance-based hard clustering algorithm, which uses distance as the criterion to determine the similarity of sample points in a dataset. That is, the smaller the distance, the higher the similarity. The algorithm considers clusters to be composed of sample points that are close together, thus making compact and independent clusters as the final result.

In a typical K-means algorithm, $K$ points in the dataset (with $n$ sample points) are selected as the initial clustering centers; Then the remaining sample points are divided into clusters with the nearest distances from the corresponding clustering centers. According to the above results, clustering centers of the new clusters are recalculated. Repeat the above steps until clustering results do not change. In this algorithm, the mean square error is generally used as a standard measure function. The $K$ generated clusters have the following characteristics: sample points in each cluster are as compact as possible, and sample points among different clusters are separated as much as possible. The time complexity of the K-means algorithm is nearly linear, and it is suitable for mining large-scale datasets.

3. The new clustering validity index

In order to effectively process overlapping dataset, we enhance the calculation of overlapping among clusters. Based on this idea, the new clustering validity index, WCH-index, is proposed. The new WCH-index can stably solve many overlapping datasets effectively. The definitions defined in the subsequent of this section are based on the following assumptions:

Suppose in Euclidean space $R^m$, a dataset $D = \{x_1, x_2, ..., x_n\}$ containing $n$ sample points is given where each sample point has $m$ attributes. The K-means algorithm divides dataset $D$ into $K$ clusters $C = \{C_1, C_2, ..., C_K\}$, and then the corresponding clustering centers are $\{v_1, v_2, ..., v_K\}$ and $V$ is the global clustering center of $D$.

**Definition 1.** The Euclidean distance between samples $x_i = \{x_{i1}, x_{i2}, ..., x_{im}\}$ and $x_j = \{x_{j1}, x_{j2}, ..., x_{jm}\}$ is defined as:

$$d(x_i, x_j) = \|x_i - x_j\| = \sqrt{(x_{i1} - x_{j1})^2 + \cdots + (x_{im} - x_{jm})^2}$$ (1)

**Definition 2.** The sum of the squared distance between clustering center point $v_i$ and the global centering point $V$ is calculated as:

$$Tr(K) = \sum_{i=1}^{K} |c_i| d(v_i, V)^2$$ (2)

**Definition 3.** The inter-cluster resolution $T_r$ is defined as:

$$T_r = \frac{Tr(K)}{K-1}$$ (3)

By calculating the Euclidean distance between the cluster center of each cluster and the global cluster center $V$, the formula can intuitively understand the degree of separation between each cluster and the global cluster center. Obviously, the higher the degree of separation is, the better the separation between clusters, and consequently, the higher the clustering quality.

**Definition 4.** In each cluster $C_i$, the sum of the squared of distance from the sample point $x_i$ to the clustering center $v_i$ is defined as ($|C_i|$ is the number of sample points in cluster $C_i$):

$$Tc(i) = \sum_{i=1}^{|C_i|} d(x_{i\mu}, v_i)^2$$ (4)

**Definition 5.** The tightness within the cluster is defined as:
\[ T_c = \frac{\sum_{i=1}^{K} T_c(i)}{(n-K)} \]  

(5)

By calculating the Euclidean distance between each sample points of a cluster to the corresponding clustering center. The degree of separation from each sample point to the corresponding clustering center can be easily understood. Obviously, the better the clustering result is, the closer the cluster internal data. That is, the smaller the compactness \( T_c(K) \) in the cluster.

**Definition 6.** In each cluster \( C_i \), the average distance between each sample point to the corresponding cluster center is calculated as:

\[ d_{ave}(C_i) = \frac{T_r(C_i)}{|C_i|} \]  

(6)

**Definition 7.** The judgment criteria for a clusters with high overlapping is defined as:

\[ \text{Term}(i,j) = d(v_i, v_j)^2 < d_{ave}(C_i) + d_{ave}(C_j) \]  

(7)

That is, calculate the Euclidean distance between the clustering centers of two clusters and the average distance between all sample points in a cluster and the corresponding clustering center. Then, count the number of clusters that satisfy the above conditions. By this, the number of clusters with high overlapping is calculated. During this process, \( \text{Term}(i,j) \) is incremented by one when the condition is met.

**Definition 8.** The overlapping between clusters \( c_i \) and \( c_j \) is defined as:

\[ T_o = \frac{\sum_{i=1, i\neq j}^{K} \text{Term}(i,j)}{n} \]  

(8)

**Definition 9.** The new cluster validity index \( \text{WCH} \) is defined as:

\[ \text{WCH}(K) = \frac{T_r}{T_c+T_o} = \frac{\sum_{i=1}^{K} T_c(i) + \sum_{i=1, i\neq j}^{K} \text{Term}(i,j)}{(n-K)} \]  

(9)

The \( \text{WCH} \) index is constructed based on features of inner-cluster tightness, inter-cluster separation and inter-cluster overlapping of dataset. By the \( \text{WCH} \) index, the separation among clusters is better we the best clustering result is acquired. That is, \( T_r \) obtains a larger value in the optimal clustering. For a cluster in the optimal clustering result, sample points are the most similar and are high compactness. That is, \( T_c \) achieves a smaller value in the optimal clustering. Meanwhile, overlapping between any two clusters is relatively small when the optimal clustering result is acquired. That is, \( T_o \) achieves a small value in the optimal clustering. In the \( \text{WCH} \) index, three features, inter-cluster separation, intra-cluster tightness and overlapping, are interactive with each other. The \( K_{opt} \) is acquired when the \( \text{WCH}(K) \) get the smallest value.

### 4. Experimental analysis

This section presents the simulation experimental results on evaluating the performance of the proposed \( \text{WCH} \) index. For computer experiments and simulations, the results were calculated using a desktop computer with Matlab 6.5, 3.4 GHz CPU, 16GB of RAM and Windows 10 Professional Version 2007. The five experimental datasets are shown in Table Irespectively.

| Dataset | Number of samples | \( K_{opt} \) | Range of \( K \) |
|---------|------------------|--------------|----------------|
| 4k2     | 400              | 4            | 2<=K<=20       |
| R1      | 200              | 5            | 2<=K<=15       |
| R15     | 600              | 15           | 2<=K<=25       |
Figure 1 shows the spatial distributions of simulated datasets 4k2 and R1. It can be clearly seen from the figure that the 4k2 dataset is divided into 4 clusters. The sample distribution has dense sample points in the same cluster, and the distribution between different clusters is relatively discrete. There is no overlapping between any of two clusters. The R1 dataset is divided into 5 clusters. The overall distribution of the dataset is relatively discrete. The distribution of sample points in each cluster is also relatively scattered. Only a small part is relatively compact. There is no obvious difference between any of two clusters and there is a certain overlapping.

Figure 2 is a diagram showing the spatial distribution of data sets R15, S1, and D31. The R15 dataset is divided into 15 clusters. The distribution of peripheral clusters satisfies the characteristics of discrete clusters and clusters within clusters. The distribution of intermediate clusters is relatively compact, and the overlap between clusters and clusters is high. The sample of the S1 dataset is divided into 15 clusters, the clusters are close to each other, and the samples in the cluster are compact. The external morphology of some clusters is quite different from other clusters, and there is a large overlap between clusters and clusters. D31 is divided into 31 clusters, the data samples are more concentrated, and the boundaries between clusters and clusters are fuzzy and overlap.

### Table 2. CVI values of 4k2 dataset.

| K  | DI  | DBI  | I   | COP | WCH |
|----|-----|------|-----|-----|-----|
| 2  | 150 | 49.25| 150 | 121.30 | 129.18 |
| 3  | 54.95 | 48.64 | 20.18 | 139.40 | 109.48 |
| 4  | 106.16 | 45.24 | 53.82 | 70.33 | 150 |
| 5  | 9.67 | 87.45 | 7.38 | 124.05 | 140.11 |
| 6  | 11.05 | 119.80 | 6.80 | 118.11 | 126.88 |
| 7  | 3.89 | 117.23 | 2.65 | 120.12 | 135.74 |
| 8  | 4.34 | 118.13 | 2.70 | 114.73 | 137.69 |
| 9  | 4.50 | 127.42 | 2.46 | 111.82 | 135.20 |
| 10 | 5.84 | 127.18 | 0.73 | 121.06 | 133.99 |
| 11 | 5.84 | 132.80 | 0.65 | 116.37 | 129.08 |
| 12 | 5.84 | 136.47 | 0.45 | 121.97 | 125.45 |
| 13 | 5.84 | 142.96 | 0.42 | 117.62 | 124.53 |
| 14 | 6.19 | 144.59 | 0.38 | 116.28 | 122.92 |
The clustering evaluation results of 4k2 dataset are shown in Table 1: The index value corresponding to the value of the optimal cluster number K is represented by a bold mark plus a descending line, that is, the optimal cluster number K opt obtained by the index. The optimal number of clusters corresponding to the 4k2 data set is 4. Therefore, it can be known that the optimal cluster number (optimal clustering) of the 4k2 dataset can be obtained through the DBI-index, the COP-index and the WCH-index; and the optimal cluster number cannot be obtained for the DI-index and the I-index (ie, The optimal number of clusters obtained is 2 and not the optimal clustering result of the 4k2 data set).

![Figure 3](image1.png)  
Figure 3. The clustering validity index values of R1 datasets. Figure 4. Cluster validity indicator values of R15 datasets. 

The optimal number of clusters for the R1 data set is 5. The five CVI is shown in Figure 3. By comparison, it is found that the CVIs WCH-index, DBI-index, COP-index, and DI-index can obtain the $K_{opt}$. However, the I-index cannot get the optimal clustering number of this dataset.

From the experimental results of 4k2 and R1, for the data sets with good separation between clusters and small overlap, WCH-index, DBI-index and COP-index can get the best cluster number, and the clustering results are stable. However, DI-index cannot obtain the optimal number of clusters, and the clustering results are unstable. I-index can't get the best number of clusters.

As shown in Figure 4, the optimal cluster number of the R15 dataset is 15. As can be seen from Figure 4, neither the I-index nor the COP-index can obtain the optimal cluster number of the dataset. The index value obtained by the I-index is far from the actual result. The results obtained by the COP-index is near to the optimal clustering number of the R15 dataset. So, this CVI can obtain the near optimal clustering number. As the optimal clustering number of R15 dataset is 15, the optimal cluster number can be obtained by the DBI-index, the DI-index and the WCH-index.

![Figure 4](image2.png)
The S1 dataset has 5000 sample points, so the clustering number K range from 2 to 70. As the special distribution shown in Fig. 2, the optimal cluster number of S1 is 15. Figure 5 shows the index values calculated by different CVIs for the S1 dataset. According to this figure 5, we can see that the DI-index, the DBI-index and the WCH-index can get the best cluster number. However, CVIs, the I-index and the COP-index cannot get the $K_{opt}$ for this dataset. Meanwhile, the performance of the I-index is the worst which only get the number 2 for the $K_{opt}$.

The D31 dataset has 31,000 sample points, so the clustering number K range from 2 to 176. Figure 2 is a sample space distribution diagram of dataset. As can be seen from the figure, the optimal cluster number of D31 is 31. For dataset D31 with high overlap, only WCH-index can obtain the best number of clusters, and other indicators cannot get the best cluster number and the result is far from the actual cluster number.

The experimental results show that WCH can effectively evaluate the tested five datasets 4k2, R1, R15, S1 and D31, and the best cluster number of them are all obtained. Since I-index needs to adjust parameters for different datasets, the optimal number of clusters for the above dataset cannot be obtained. The DI-index, DBI-index and COP-index indicators are more stable for datasets without overlapping data, and basically the best number of clusters can be obtained. For a data set with a small degree of overlap, the COP-index has been unable to obtain the optimal number of clusters. The DI-index and DBI-index can still achieve the best number of clusters. However, when the dataset overlap is high, DI-index and DBI-index cannot correctly evaluate the clustering results, and the optimal number of clusters cannot be obtained. For the above data sets, regardless of whether there is overlapping data, the data set overlaps, the WCH-index can accurately evaluate the clustering results and obtain the optimal number of clusters. Experiments have demonstrated the stability of WCH-index and the accuracy of data sets with high overlap.

5. Conclusion and outlook
This paper proposed a new cluster validity index (WCH) to calculate the optimal clustering number for overlapping data sets. The WCH-index consisted of three factors, the inner-cluster tightness, the inter-cluster resolution, and inter-cluster overlapping. Through the linear combination of the three factors, and their mutual constraints, the WCH index can effectively process datasets with high overlapping and accurately find the optimal clustering numbers for many data set. However, the WCH-index is still insufficiency in dealing with large scale datasets. So, in the future, we will improve this shortcoming.

References
[1] J Han, M Kamber. Data Mining: Concepts and Techniques. Data Mining Concepts Models Methods & Algorithms Second Edition, 5(4), 2011, pp.1–18.
[2] Omran M G H, Engelbrecht A P, Salman A. An overview of clustering methods. Intelligent Data Analysis, 2007, 11(6), pp. 586-605.

[3] Rathore P, Ghafoori Z, Bezdek J C, et al. Approximating Dunn’s Cluster Validity Indices for Partitions of Big Data[J]. IEEE Transactions on Cybernetics, 2018, PP(99).

[4] JC Dunn. A fuzzy relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. Journal of Cybernetics, 1974, 3(3), pp. 32-57.

[5] DL Davies, DW Bouldin. A Cluster Separation Measure. IEEE Transactions on Pattern Analysis & Machine Intelligence, 1979, PAMI-1 (2), pp. 224-227.

[6] S Bandyopadhyay, U Maulik. Nonparametric genetic clustering: comparison of validity indices. IEEE Transactions on Systems Man & Cybernetics Part C Applications & Reviews, 2001, 31(1), pp. 120-125.

[7] I Gurrutxaga, I Albisua, O Arbelaitz, Mart, N Jos. SEP/COP: An efficient method to find the best partition in hierarchical clustering based on a new cluster validity index. Pattern Recognition, 2010, 43 (10), pp. 3364-3373.