ISSN 1330-3651 (Print), ISSN 1848-6339 (Online) https://doi.org/10.17559/TV-20200828055024
Original scientific paper

Understanding the Evaluation Abilities of External Cluster Validity Indices to Internal Ones

Xiaonan GAO, Guiying WEI, Sen WU*, Falong FAN

Abstract: Evaluating internal Cluster Validity Index (CVI) is a critical task in clustering research. Existing studies mainly employ the number of clusters (NC-based method) or external CVIs (external CVIs-based method) to evaluate internal CVIs, which are not always reasonable in all scenarios. Additionally, there is no guideline of choosing appropriate methods to evaluate internal CVIs in different cases. In this paper, we focus on the evaluation abilities of external CVIs to internal CVIs, and propose a novel approach, named external CVI's evaluation Ability MEasurement approach through Ranking consistency (CAMER), to measure the evaluation abilities of external CVIs quantitatively, for assisting in selecting appropriate external CVIs to evaluate internal CVIs. Specifically, we formulate the evaluation ability measurement problem as a ranking consistency task, by measuring the consistency between the evaluation results of external CVIs to internal CVIs and the ground truth performance of internal CVIs. Then, the superiority of CAMER is validated through a real-world case. Moreover, the evaluation abilities of seven popular external CVIs to internal CVIs in six different scenarios are explored by CAMER. Finally, these explored evaluation abilities are validated on four real-world datasets, demonstrating the effectiveness of CAMER.

Keywords: cluster validity indices (CVIs); evaluation ability; quantitative measurement; ranking consistency

1 INTRODUCTION

Clustering is to discover unknown clusters in mass data, which directly depends on some prior choices, such as clustering algorithm, similarity metric and parameter configuration, etc [1, 2]. For the same dataset, different algorithms, similarity metrics or parameter configurations would produce various clustering results. Thus, Cluster Validity Indices (CVIs) are demanded to estimate which one is the best. Existing studies on CVIs can be divided into external CVIs [3] and internal CVIs [4]. The difference between them is whether the data labels or other external information are required. External CVIs evaluate clustering results by assessing the consistency between the clustering label and the ground truth label. Internal CVIs do evaluation according to the compactness of intra-cluster and separation of inter-clusters.

Since most of datasets in real-world lack labels or other external information, internal CVIs are more practical [5]. Along this line, evaluating internal CVIs becomes a critical and challenging research problem. Take researchers studying internal CVIs for example, if a researcher proposes an internal CVI, an evaluation method is required to assess the effectiveness of the newly presented internal CVI. Accurate evaluation will guide others to study or use this CVI in the right direction, otherwise, it may result in misguidance. In this paper, we aim to address the problem of precise evaluation of internal CVIs, by investigating the evaluation abilities of existing methods to internal CVIs in different scenarios.

There are mainly two ways to solve the problem in existing studies: (1) An internal CVI would be considered to be good, if it could identify the correct number of clusters *nc*, we call it NC-based method [5, 6]. However, there may be a partition P whose number of clusters is not equal to *nc* that fits to the dataset better than the partition P^{nc} whose number of clusters is *nc* [7]. Hence, it is not always reasonable to exploit NC-based method to evaluate internal CVIs. (2) External CVIs are utilized to evaluate internal CVIs by measuring the consistency between clustering partitions with the best results of internal CVIs and ground truth partition, called external CVIs-based method [4]. This method does not rely on the number of clusters, but the evaluation result is affected severely by the external CVI used. Different external CVIs might produce various evaluation results for the same internal CVI, since they have some biased behaviours [3]. For instance, some external CVIs show a monotonous bias for the number of clusters [8-11], and some of them favour the balanced clustering results even if the dataset is skewed [12]. Consequently, the existing two methods are not always feasible to evaluate internal CVIs in different scenarios. To this end, how to choose an appropriate method to evaluate internal CVIs in different cases is an urgent task to be solved. In this paper, we focus on the external CVIs-based method and try to figure out the evaluation abilities of several well-known external CVIs to internal CVIs in different situations, in order to assist in selecting an appropriate external CVI to evaluate internal CVIs in a certain case.

To the best of our knowledge, this may be the first attempt to explore the evaluation abilities of external CVIs to internal CVIs quantitatively. Existing relevant researches on external CVIs [3] mainly focus on assessing their evaluation abilities to clustering partitions directly, not to internal CVIs that we care about in this work. Moreover, most of them employ the number of clusters to indicate the effectiveness of external CVIs, leading to that the final evaluation conclusions are affected severely by the correct number of clusters *nc*.

In this paper, we propose a novel approach, named external CVI's evaluation Ability MEasurement approach through Ranking consistency (CAMER), by formulating the external CVI's evaluation ability measurement problem as a ranking consistency task.This approach can assess the evaluation abilities of external CVIs to internal CVIs quantitatively and help choose reasonable external CVIs to evaluate in different scenarios. Specifically, given multiple clustering partitions with known goodness, several internal CVIs are used to evaluate these partitions, and we can rank these internal CVIs based on the goodness of their identified partitions. This ranking is regarded as the ground truth ranking. Then, an external CVI evaluates the partitions with the best results of internal CVIs, and the evaluation ranking can be obtained based on the external CVI's scores. Moreover, two popular ranking correlation

measurements, Kendall's tau coefficient [13] and Spearman's footrule [14] are utilized to measure the consistency between the ground truth ranking and the evaluation ranking. The more consistent the two rankings, the more accurate the evaluation ability of the external CVI to internal CVIs. The basic idea is that the ground truth ranking represents the actual differences of multiple internal CVIs, if the evaluation ranking produced by an external CVI is consistent with the ground truth ranking, which means the external CVI is able to distinguish the differences. Furthermore, the superiority of CAMER is illustrated based on a real-world case. And we explore the evaluation abilities of seven well-known external CVIs to internal CVIs in six different scenarios: spherical distribution, density, irregular shape, noise, skewed distribution, and subclusters by using CAMER, which can be used to assist in selecting suitable external CVIs to evaluate internal CVIs. Finally, we validate the effectiveness of the explored evaluation abilities via four real-world datasets.

The main contributions of our work are summarized as follows:(1) A novel approach, named CAMER, is proposed to measure the evaluation abilities of external CVIs to internal CVIs quantitatively (Section3); (2) The superiority of CAMER is interpreted based on a real-world case (Section4); (3) The evaluation abilities of seven popular external CVIs to internal CVIs in six different scenarios are summarized by CAMER (Section5); (4) The evaluation abilities summarized by CAMER are verified through four real-world datasets with different structures (Section 6).

2 RELATED WORK

In this section, we review the internal CVI evaluation methods and summary of several relevant studies in recent years in Tab. 1, mainly classified from two perspectives: NC-based method and external CVIs-based method.

Table 1 Summary of several relevant researches on internal CVI evaluation

Study	NC-based	External CVIs-based
Xie et al. (2020) [15]	√	√
Hu et al. (2019) [16]	╳	√
Gao and Wu (2019) [17]	╳	√
Fu et al. (2019) [18]	╳	√
Cheng et al.(2018) [19]	√	√
Gao and Yang (2018) [20]	╳	√
Kim et al. (2018) [21]	╳	√
Zhou et al. (2018) [22]	√	╳
Zhao et al. (2017) [23]	╳	√
Thomas et al. (2017) [24]	√	╳
Zhou et al. (2016) [25]	√	╳
Fu et al. (2016) [26]	√	√

2.1 External CVIs-Based Method

External CVIs are the most common way to evaluate internal CVIs. The basic assumption is that, given multiple clustering partitions of a dataset, an internal CVI evaluates these partitions respectively, the consistency between the best clustering partition and the ground truth partition is measured by external CVI. The more consistent, the better the evaluation performance of the internal CVI. As shown in Tab. 1, most studies utilize external CVIs-based method to do evaluation. Xie et al. only used F-measure (F) to evaluate their presented internal CVI; Hu et al. utilized cluster-level Centroid Index (CI) and Purity (P) in their study; Seven external CVIs, including Accuracy (A), Adjusted Rand Index (ARI), F, Micro-p (M), Normalized Mutual Information (NMI), P and Rand Index (RI), were exploited in the study of Gao and Wu; Fu et al. utilized NMI to do evaluation; Cheng et al. did evaluation in their experiments through A; Gao and Yang evaluated a new internal CVI by ARI and NMI; A, Balanced Accuracy (BA) and Balanced Correction Rate (BCR) were exploited in Kim et al.'s research; Zhao et al. employed A, ARI and NMI to do evaluation; And Fu et al. evaluated the internal CVIs through NMI and A. It is easy to notice that most of these studies only employed three or fewer external CVIs to evaluate, except Gao and Wu's research in which seven external CVIs were utilized.

However, the evaluation abilities of external CVIs to internal CVIs are unstable, since there are biased behaviours of external CVIs, such as monotonous bias. Therefore, the convincing evaluation result cannot be obtained by only using a few external CVIs. It is a solution to use as many external CVIs as possible, but it is time consuming and difficult to conduct analysis if the results do not converge. In addition, there lacks a guideline of choosing suitable external CVIs to evaluate internal CVIs in different cases. In this paper, we try to address this problem, by proposing a novel approach to measure the evaluation abilities of external CVIs to internal CVIs quantitatively.

2.2 NC-Based Method

The basic idea of NC-based method is whether an internal CVI can identify the clustering partition with the correct number of clusters. If the number of clusters hit by an internal CVI is correct, we will consider this CVI exhibits good performance.

In recent studies as listed in Tab. 1, we can see that there are six researches that employed NC-based method to do evaluation, and three of them used external CVIs-based method simultaneously. In brief, three of the 12 studies utilized only NC-based method to evaluate internal CVIs, thus we can know that most researches on internal CVIs no longer exploited NC-based method alone to do evaluation. This is because the underlying assumption of NC-based method is that the number of clusters of the best partition is equal to the real number of clusters *nc*. Nevertheless, this assumption does not always hold [7]. Therefore, NC-based method cannot evaluate internal CVIs reasonably in all scenarios.

Besides, NC-based method is also often used to evaluate external CVIs. We have found that most of the related works focused on validating the performance of external CVIs to clustering partitions directly, not the evaluation abilities of them to internal CVIs that we aim to study in this paper. However, we believe that NC-based method can be transferred to solve the evaluation ability measurement problem of external CVIs to internal CVIs, through the following calculation process: (1) Calculating evaluation scores: Calculate the evaluation scores of an external CVI to multiple internal CVIs. Given several clustering partitions for the same dataset, internal CVIs

evaluate these partitions and hit the best partitions respectively. Then, we calculate the evaluation scores of an external CVI on the partitions hit by internal CVIs. (2) Finding the optimal evaluation score: After obtaining the evaluation scores of the external CVI, we pick out the partition with the optimal score, and then identify its number of clusters. (3) Comparing with the correct number of clusters: The number of clusters of the identified partition is compared with the correct number of clusters. If they are equal, we consider that the external CVI exhibits excellent evaluation ability to internal CVIs, otherwise, its evaluation ability is poor. There are two limitations of NC-based method, one is that NC-based method can only produce two qualitative conclusions, namely good or bad, since the number of clusters is the only evaluation criterion. Furthermore, if there is no partition with the correct number of clusters, NC-based method will fail, which can output only negative conclusion.

To this end, we propose a new approach CAMER in Section 3, to measure the evaluation abilities of external CVIs to internal CVIs quantitatively, which is capable of overcoming the limitations of traditional NC-based method and finding a new way to tackle the evaluation ability measurement task.

3 CAMER

In this section, we propose a new approach, named external CVI's evaluation Ability MEasurement approach through Ranking consistency (CAMER), to assess the evaluation abilities of external CVIs to internal CVIs quantitatively.

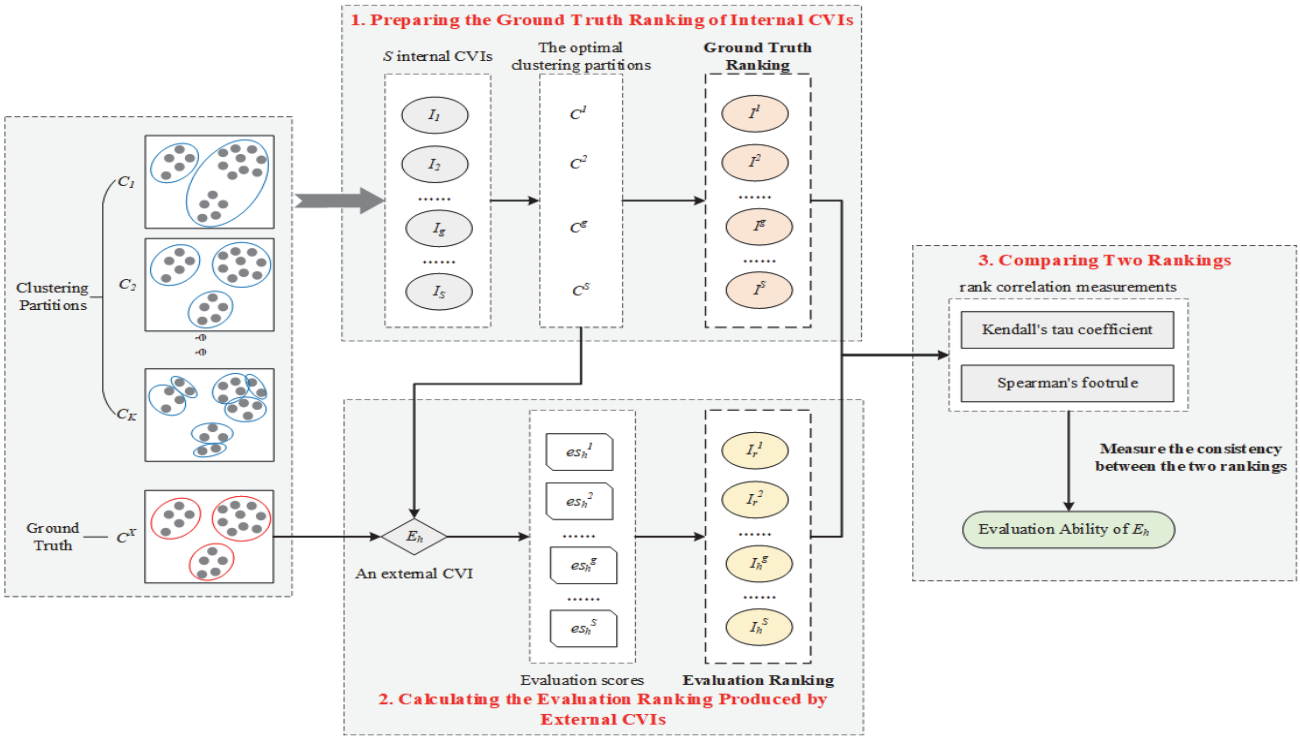

Figure 1 Overview of CAMER

3.1 Overview of CAMER

The basic idea of CAMER is to measure the consistency between the evaluation results of external CVIs to internal CVIs and the ground truth performance of internal CVIs. As shown in Fig. 1, CAMER contains three steps. (1) Preparing the ground truth ranking: Given a set of clustering partitions with known goodness, several internal CVIs evaluate these partitions and identify the optimal one. Along this line, the ground truth ranking of internal CVIs can be obtained based on the goodness of the corresponding optimal clustering partitions, which reveals the actual differences of these internal CVIs. (2) Calculating the evaluation ranking: An external CVI is used to evaluate the performance of the optimal clustering results identified by internal CVIs, and we can get the evaluation ranking of internal CVIs according to external CVI's evaluation scores, which represents the differences of internal CVIs recognized by the external CVI. (3) Comparing two rankings: We exploit two rank correlation measurements, Kendall's tau coefficient and Spearman's foot rule, to measure the consistency between the two rankings quantitatively, if they are consistent, that means the external CVI is able to distinguish the actual differences correctly, in other words, it can evaluate internal CVIs accurately. The more consistent the two rankings are, the more accurate the evaluation ability of the external CVI to internal CVIs is.

3.2 Preparing the Ground Truth Ranking

Given a dataset X, its ground truth partition is C^X. There are K clustering partitions $C = \{C_1, C_2, \ldots, C_j, \ldots, C_K\}$, for X, and the goodness of each clustering partition C_j, $1 \le j \le K$ is known. S internal CVIs, $\{I_1, I_2, \ldots, I_g, \ldots, I_S\}$, evaluate the K clustering partitions respectively and identify the corresponding optimal clustering partitions $C_S^{opt} = \left\{C^1, C^2, \ldots, C^g, \ldots, C^S\right\}$, based on their evaluation results, where C^g infers the optimal clustering partition

identified by I_g. Since we know the goodness of each clustering partition, we can rank the S internal CVIs as:

$$IRank = \left[I^1, \mathrm{I}^2, \ldots, I^g, \ldots, I^S \right] \quad (1)$$

Where I^1 represents the internal CVI on this position ranking first and I^g is the internal CVI ranking gth. The higher the ranking of internal CVI, the better the clustering partition it identifies. *IRank* is regarded as the ground truth ranking of internal CVIs for dataset X.

Taking C_1 and C_2 in Fig. 1 as an example, we already know C_2 fits dataset X better than C_1. There are two internal CVIs I_1 and I_2, they evaluate C_1 and C_2 respectively, and I_1 identifies C_1 is better than C_2, conversely, the evaluation result of I_2 on C_2 outperforms that on C_1. Therefore, we obtain the optimal clustering partitions identified by I_1 and $I_2 : C_S^{opt} = \left\{ C^1, C^2 \right\}$. And then, the ground truth ranking of these two internal CVIs can be prepared: $IRank = \left[I_2, \mathrm{I}_1 \right]$, according to the goodness of C_1 and C_2.

3.3 Calculating the Evaluation Ranking

There are R external CVIs $\left\{ E_1, E_2, \ldots, E_h, \ldots, E_R \right\}$, we utilize each external CVI, E_h, $1 \le h \le R$, to evaluate the optimal clustering partitions C_S^{opt} identified by S internal CVIs, and get the evaluation scores $ES_h\left(C_S^{opt}\right) = \left\{ es_h^1, es_h^2, ..., es_h^g, ..., es_h^s \right\}$. According to the scores, we can rank the S internal CVIs as:

$$IRank_h = \left\{ I_h^1, I_h^2, ..., I_h^g, ..., I_h^s \right\} \quad (2)$$

Where I_h^1 represents E_h regards the performance of the internal CVI on this position is the best. The higher the ranking of internal CVI, the better the evaluation result of external CVI on internal CVI. $IRank_h$ is the evaluation ranking of S internal CVIs produced by external CVI E_h. Considering the example in the previous subsection, for internal CVIs I_1 and I_2, their optimal clustering partitions are C_1 and C_2 respectively. An external CVI E is used to evaluate C_1 and C_2 and obtains two evaluation scores es^1 and es^2. If es^1 is better than es^2, we can claim that I_1 identify a better clustering partition than I_2, from the perspective of E. To this end, the evaluation ranking of I_1 and I_2 produced by E is $IRank_E = [I_1, I_2]$. Otherwise, if es^2 is better than es^1, the evaluation ranking will be $IRank_E = [I_2, I_1]$.

3.4 Comparing Two Rankings

After getting the ground truth ranking *IRank* and the evaluation ranking $IRank_h$, we use two well-known rank correlation measurements Kendall's tau coefficient and Spearman's footrule to measure the consistency between these two rankings. They can assess the similarity of the orderings of two ranking lists and their ranges of values are [0, 1]. The larger values of Kendall and Spearman mean the two rankings are more consistent, that means the evaluation result of E_h to S internal CVIs is close to the ground truth. Along this line, we can say that the evaluation ability of E_h to internal CVIs is accurate. Otherwise, when the ground truth ranking *IRank* and the evaluation ranking $IRank_h$ are dissimilar, the values of Kendall and Spearman will be small, indicating the evaluation ability of E_h to internal CVIs is weak.

For example, we have prepared the ground truth ranking $IRank = [I_2, I_1]$, if the evaluation ranking produced by E is $IRank_E = [I_1, I_2]$, these two rankings are not consistent, and the values of Kendall and Spearman would be small. If the evaluation ranking is $IRank_E = [I_2, I_2]$, which is the same as the ground truth ranking, the Kendall value and Spearman value would be equal to 1.

CAMER formulates the external CVI's evaluation ability measurement problem as a ranking consistency task, which realizes the quantitative assessment of the evaluation abilities of external CVIs to internal CVIs.

4 COMPARISON OF CAMER AND NC-BASED METHOD

In this section, we demonstrate the superiority of CAMER, by comparing it with existing NC-based method through a real-world case. We first introduce the case dataset used. Then, NC-based method and CAMER are employed to measure the evaluation abilities of seven popular external CVIs [17] to internal CVIs respectively. Finally, we summarize the differences between the two methods, to clarify the superiority of CAMER. The comparison is conducted in an environment of I5-7300HQCPU @ 2.50GHz, MATLAB R2018b.

4.1 Data Collection and Preparation

The dataset in this real-world case is about the locations of education and training institutions in Beijing, collected from a POI data open platform of China (http://www.poilist.cn/). There are 820 education and training institutions, of which each one is described by the longitude and latitude of its location, and the district where it is located. In this case, the districts include Yanqing, Huairou, Miyun, Pinggu, Mentougou and Other Districts (other districts in Beijing except the aforementioned five districts). As shown in Fig. 2, we regard the districts as the labels; in this way, the dataset is partitioned into six classes, indicated by different colours. Notably, the correct number of clusters of this dataset is six. In addition, nine clustering partitions are prepared by K-means algorithm with the number of clusters from 2 to 10 (NC = 2 to NC = 10), shown in Fig. 3.

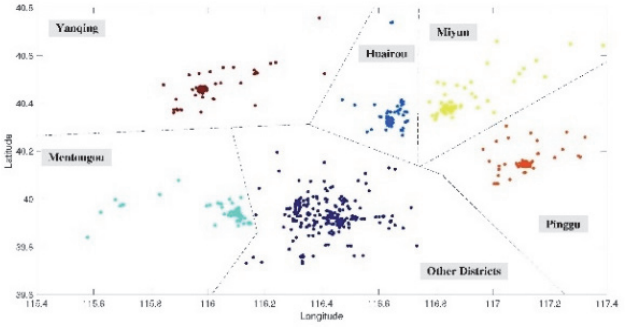

Figure 2 Visualization of dataset in the real-world case

4.2 Comparison Results

In this subsection, we report the comparison results of CAMER and NC-based method. To illustrate the effect of the correct number of clusters on the measurement result, we conduct comparison based on two kinds of partitions: (1) Partitions including the correct number of clusters; (2) Partitions excluding the correct number of clusters.

4.2.1 Partitions Including the Correct Number of Clusters

We do comparison based on the nine partitions including NC = 6 (the correct number of clusters) in Fig. 3. The upper part of Tab. 2 lists the evaluation ability measurement results produced by NC-based method. We already know that the correct number of clusters is six. To this end, we can conclude that ARI, F, NMI, and RI have good evaluation abilities to internal CVIs, since the number of clusters of the partition identified by them is equal to the correct number of clusters. Nevertheless, the evaluation abilities of A, M, and P are weak. Obviously, there are only two qualitative conclusions produced by NC-based method, namely good "√" and bad "╳".

The measurement results on partitions including the correct number of clusters by CAMER are shown in the left part of Tab. 3. It is clear that the Kendall and Spearman values of ARI, F, NMI and RI are all ones, so we can say that these four external CVIs can evaluate internal CVIs accurately. Conversely, the evaluation abilities of A, M and P to internal CVIs are poor. It is noteworthy that the evaluation conclusions produced by CAMER are specific values. Based on this, we can compare the evaluation abilities of different external CVIs quantitatively, rather than only draw qualitative conclusions, such as good or bad. Furthermore, according to the evaluation results output by NC-based method and CAMER, we find their conclusions are consistent, the evaluation abilities of ARI, F, NMI and RI outperform that of A, M and RI. Therefore, we know that the evaluation results of CAMER are precise.

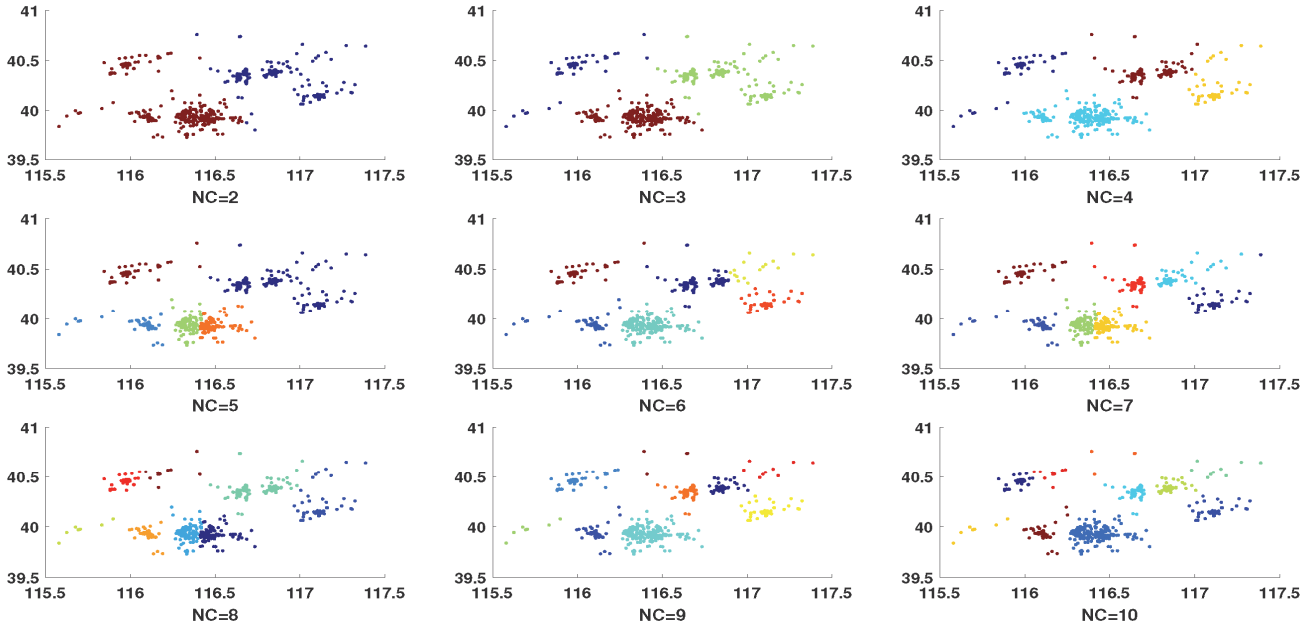

Figure 3 Clustering partitions of the case dataset

4.2.2 Partitions Excluding the Correct Number of Cluster

Here, we compare CAMER and NC-based method based on the eight partitions (NC = 2 to NC = 5, NC = 7 to NC = 10) in Fig. 3, except the partition with six number of clusters (NC = 6). The measurement results by NC-based method are reported in the lower part of Tab. 2. It is worth to note that the number of clusters of the identified partitions is not equal to the correct number of clusters. In this way, NC-based method concludes that these seven external CVIs show poor evaluation abilities to internal CVIs. Apparently, when there is no partition with the correct number of clusters in the given partitions, NC-based method will fail, which will come to the negative conclusions for all external CVIs.

The evaluation abilities on partitions excluding the correct number of clusters measured by CAMER are shown in the right part of Tab. 3. It is obvious that ARI, F, NMI and RI are capable of assessing internal CVIs accurately. But the performance of A, M and P is poor. If we compare the left and right parts of Tab. 3, it should be noticed that the two measurement results are consistent, even if the partition with the correct number of clusters (NC = 6) does not appear in the right part, which means our CAMER is not affected by the correct number of clusters as NC-based method. It can still produce accurate evaluation conclusions, when the partition with the correct number of clusters is not given.

Table 2 Measurement results on the case dataset by NC-based method

	Partitions including the correct number of clusters						
External CVI	A	ARI	F	M	NMI	P	RI
Identified partition	NC = 2	NC = 6	NC = 6	NC = 2	NC = 6	NC = 2	NC = 6
Conclusion	╳	√	√	╳	√	╳	√
	Partitions excluding the correct number of clusters						
External CVI	A	ARI	F	M	NMI	P	RI
Identified partition	NC = 2	NC = 7	NC = 7	NC = 2	NC = 7	NC = 2	NC = 7
Conclusion	╳	╳	╳	╳	╳	╳	╳

Table 3 Measurement results on the case dataset by CAMER

	Partitions including the correct number of clusters		Partitions excluding the correct number of clusters	
	Kendall	Spearman	Kendall	Spearman
A	−0.2424	−0.3147	−0.3939	-−0.4685
ARI	1	1	1	1
F	1	1	1	1
M	−0.2424	−0.3147	−0.3939	−0.4685
NMI	1	1	1	1
P	−0.2424	−0.3147	−0.3939	−0.4685
RI	1	1	1	1

4.3 Difference Summary

Based on the evaluation results of CAMER and NC-based method, we can summarize their differences:

(1) In terms of the form of evaluation result, existing NC-based method can only output qualitative conclusion, namely good ability or bad ability. Our proposed CAMER can produce quantitative conclusion, in this way, the gap of evaluation abilities between different external CVIs can be measured accurately.

(2) For the effect of the correct number of clusters, NC-based method is affected by the correct number of clusters severely, which will be invalid if there is no partition with the correct number of clusters in the given partitions. CAMER is not affected by the correct number of clusters, which is still able to measure the evaluation abilities of external CVIs precisely.

Overall, our proposed CAMER approach realizes more precise and stable measurement for the evaluation abilities of external CVIs to internal CVIs than existing method. Hereafter, we will explore the evaluation abilities of seven popular external CVIs to internal CVIs based on six different structure datasets by employing CAMER, for guiding the selection of external CVIs to evaluate.

5 EVALUATION ABILITY EXPLORATION

In this section, we generate six synthetic datasets with different structures to explore the evaluation abilities of seven well-known external CVIs to internal CVIs by CAMER in different cases: spherical distribution, density, irregular shape, noise, skewed distribution and subclusters. The exploration can guide the selection of appropriate external CVIs to evaluate internal CVIs in a certain case.

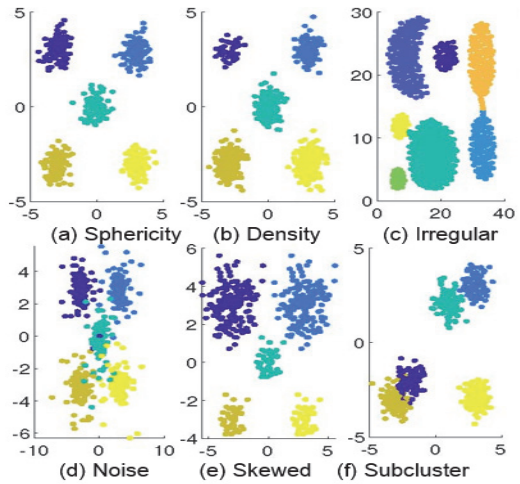

Figure 4 Visualization of six different structure datasets

5.1 Synthetic Dataset Preparation

We generate a dataset with five spherical clusters as shown in Fig. 4a. Based on this spherical dataset, a dataset with different cluster densities is generated, as shown in Fig. 4b. There are five spherical clusters, their radii are the same, but the number of instances in different clusters is distinct. Hence, the cluster densities in this dataset are various. Next, the irregular shaped dataset Aggregation [27] is exploited directly as shown in Fig. 4c. Additionally, based on the spherical dataset, we add some noises to each cluster and generate a noised dataset as shown in Fig. 4d. Moreover, the skewed dataset with unbalanced distribution is generated as shown in Fig. 4e. And Fig. 4f shows the dataset with two subclusters.

For the six synthetic datasets, each of them is clustered into nine clustering partitions by K-means algorithm with the number of clusters from 2 to 10. Based on this, we use CAMER to explore the evaluation abilities of seven well-known external CVIs to internal CVIs.

5.2 Evaluation Ability Exploration on Synthetic Datasets

Tab. 4 reports the evaluation ability measurement results of the seven external CVIs to internal CVIs by CAMER on the six different structure datasets. As for spherical dataset, we can see that all Kendall values are greater than 0.84, and all Spearman values are greater than 0.93, which means the seven external CVIs have good evaluation abilities to internal CVIs on spherical dataset. In terms of dataset with different cluster densities, it is clear that the Kendall values and Spearman values are all ones. Therefore, the seven external CVIs can assess internal CVIs accurately on dataset with different cluster densities. For irregular shaped dataset, the Kendall values and Spearman values of ARI, F and RI are all ones, indicating these three external CVIs can handle irregular shaped dataset. However, the results of A, M, NMI and P are less than 0.5, thus, these four external CVIs are not suitable to evaluate internal CVIs on this irregular shaped dataset. As to noised dataset, it is obvious that the Kendall values and Spearman values of A, ARI, F, NMI, P and RI are all greater than 0.96, which means these six external CVIs can evaluate internal CVIs on noised dataset effectively. Nevertheless, M performs poorly. In terms of skewed dataset, we can see that the Kendall values and Spearman values of A, ARI, F, M and P are all one, demonstrating these five external CVIs are able to evaluate internal CVIs effectively on skewed dataset. Conversely, NMI and RI cannot assess internal CVIs accurately in this case. Finally, for dataset with subclusters, it is worth to note that the Kendall values and Spearman values of ARI, F, M, NMI and RI are all ones, that means the five external CVIs can assess internal CVIs precisely. Additionally, the measurement results of A and P are less than 0.3, implying they are not suitable to evaluate internal CVIs in this case.

In brief, we note that different external CVIs show various evaluation performances. Thus, it is valuable to summarize the evaluation abilities of external CVIs, for guiding the selection of appropriate external CVIs to evaluate internal CVIs in a certain case.

Table 4 Measurement results on the six different structure datasets by CAMER

	Sphericity		Density		Irregular	
	Kendall	Spearman	Kendall	Spearman	Kendall	Spearman
A	0.8485	0.9371	1	1	0.3939	0.4825
ARI	1	1	1	1	1	1
F	1	1	1	1	1	1
M	0.8485	0.9371	1	1	0.4545	0.4545
NMI	1	1	1	1	0.3939	0.3636
P	0.8485	0.9371	1	1	0.3939	0.4825
RI	1	1	1	1	1	1
	Noise		Skewed		Subcluster	
	Kendall	Spearman	Kendall	Spearman	Kendall	Spearman
A	0.9697	0.9930	1	1	0.2424	0.2937
ARI	1	1	1	1	1	1
F	1	1	1	1	1	1
M	−0.0606	−0.0839	1	1	1	1
NMI	1	1	0.3939	0.3497	1	1
P	0.9697	0.9930	1	1	0.2424	0.2937
RI	1	1	0.3939	0.3497	1	1

Table 5 Summary of evaluation abilities of seven external CVIs

	A	ARI	F	M	NMI	P	RI
Sphericity	√	√	√	√	√	√	√
Density	√	√	√	√	√	√	√
Irregular	-	√	√	-	-	-	√
Noise	√	√	√	-	√	√	√
Skewed	√	√	√	√	-	√	-
Subcluster	-	√	√	√	√	-	√

5.3 Evaluation Ability Summary

We summarize the evaluation abilities of the seven well known external CVIs to internal CVIs in different scenarios derived from our experiments in Tab. 5, which can assist in the selection of external CVIs conveniently and reasonably to evaluate internal CVIs:

(1) For spherical dataset, the seven external CVIs can evaluate internal CVIs accurately;

(2) For dataset with different cluster densities, the seven external CVIs show good performance;

(3) For irregular shaped dataset, ARI, F and RI can be used to evaluate internal CVIs;

(4) For noised dataset, A, ARI, F, NMI, P and RI are able to evaluate internal CVIs reasonably;

(5) For skewed dataset, A, ARI, F, M and P exhibit superior evaluation abilities than other two CVIs;

(6) For dataset with subclusters, ARI, F, M, NMI and RI perform well in the evaluation of internal CVIs.

6 EVALUATION ABILITY VALIDATION

In this section, four real-world datasets with different structures are exploited to validate the evaluation ability summary concluded by CAMER in Section 5.

6.1 Real-World Dataset Collection

Tab. 6 describes these datasets, whose characteristics are shown in the last column. It is easy to notice that the characteristics of these datasets cover all the data structures studied in the last section, so that we believe the validation on these datasets is convincing.

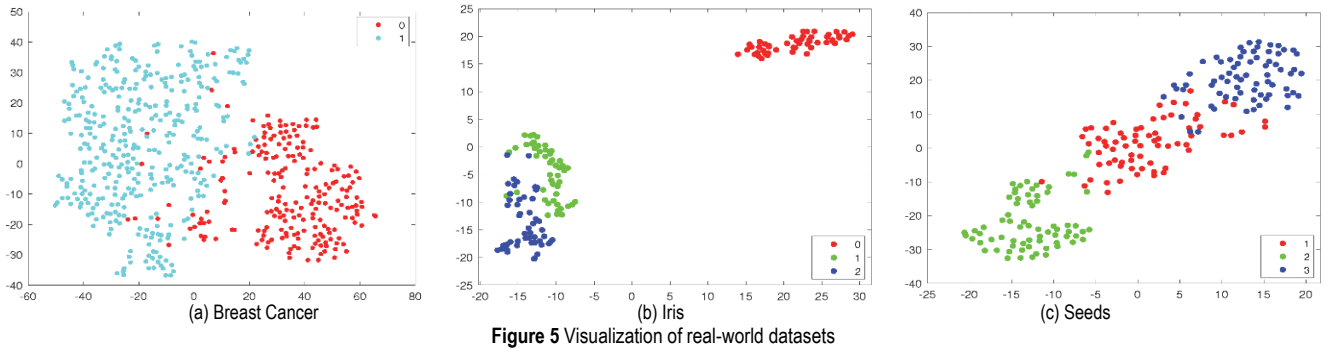

Figure 5 Visualization of real-world datasets

Specifically, POI dataset is derived from the real-world case in Section 4, and the other three datasets are collected from the UCI machine learning repository [28]. Moreover, the structure of each dataset is summed up based on its visualization. We use t-Distributed Stochastic Neighbour Embedding [29] (t-SNE) to reduce the dimensions of the three UCI datasets for visualization as shown in Fig. 5, and then the characteristics can be analysed manually. We employ K-means algorithm to cluster these datasets into multiple partitions with different number of clusters from 2 to 20. Based on this, the evaluation abilities of the aforementioned seven external CVIs to internal CVIs are measured by CAMER, to validate our summary in the last section.

6.2 Evaluation Ability Validation on Real-World Datasets

Tab. 7 reports the evaluation ability measurement results on four real-world datasets. For POI dataset. We note that ARI, F, NMI and RI exhibit excellent evaluation abilities to internal CVIs. On the contrary, the rest three CVIs, namely A, M and P, show relatively weak performance. This is in keeping with our summary in Tab. 5, that ARI, F, NMI and RI are the only four external CVIs which can evaluate internal CVIs precisely on dataset with noise and subclusters simultaneously. As to Breast Cancer dataset, we notice that ARI, F, NMI and RI show excellent evaluation abilities to internal CVIs. Conversely, the evaluation abilities of A, M and P are relatively poor. Notably, while Breast Cancer is skewed, NMI and RI perform well, that is not exactly the same as the summary in Tab. 5. The possible reason may be the other two characteristics "Density" and "Subcluster" are dominated in this dataset. Along this line, we believe the conclusions in Breast Cancer are consistent with our summary overall. As for Iris dataset, the performance of ARI, F and RI outperform that of other four CVIs significantly, that is consistent with our summary about these three external CVIs. ARI, F and RI can deal with the dataset with subclusters, noised samples and irregular shape effectively.

Finally, in terms of Seeds dataset, ARI, F, M and RI can evaluate internal CVIs accurately, and NMI exhibits the second-best performance, moreover, the remaining two CVIs, A and P, perform poor. The results are in keeping with our summary, that ARI, F, M and RI are able to process the spherical dataset with subclusters.

Table 6 Summary of four real-world datasets for evaluation ability validation

Name	# Instances	# Attributes	# Classes	Characteristics
POI	820	2	6	Noise, Subcluster
Breast Cancer	569	30	2	Density, Skewed, Subcluster
Iris	150	4	3	Irregular, Noise, Subcluster
Seeds	210	7	3	Sphericity, Subcluster

Consequently, in this section, the measurement results on four real-world datasets with different characteristics validate the effectiveness of our evaluation ability summary for seven popular external CVIs in Tab. 5, demonstrating these explored evaluation abilities can be employed to guide the selection of appropriate external CVIs for the evaluation of internal CVIs.

Table 7 Measurement results on four real-world datasets by CAMER

	POI		Breast Cancer	
	Kendall	Spearman	Kendall	Spearman
A	0.4242	0.5455	0.6667	0.7483
ARI	1	1	1	1
F	1	1	1	1
M	0.4242	0.5455	−0.0606	−0.014
NMI	1	1	1	1
P	0.4242	0.5455	0.6667	0.7482
RI	1	1	1	1
	Iris		Seeds	
	Kendall	Spearman	Kendall	Spearman
A	0.1212	0.1469	−0.1515	−0.2168
ARI	1	1	1	1
F	1	1	1	1
M	0.1212	0.1469	1	1
NMI	0.1212	0.1469	0.6667	0.7483
P	0.1212	0.1469	−0.1515	−0.2169
RI	1	1	1	1

7 CONCLUSIONS

In this paper, we investigate how to choose an appropriate method to evaluate internal CVIs. Along this line, we propose a new approach, named external CVI's evaluation Ability MEasurement approach through Ranking consistency (CAMER), to measure the evaluation abilities of external CVIs to internal CVIs quantitatively, by formulating this problem as a ranking consistency task. Specifically, we first prepare the ground truth ranking of several internal CVIs based on goodness of clustering partitions. Then, the evaluation ranking of internal CVIs is obtained according to the scores of external CVI. Finally, the consistency between the two rankings is calculated by using rank correlation measurements, which can reflect the evaluation abilities of external CVIs to internal CVIs quantitatively. The superiority of CAMER is interpreted by comparing with existing NC-based method on a real-world case. And the evaluation abilities of seven well-known external CVIs to internal CVIs in six different scenarios are explored and summarized, which can be used to assist in choosing suitable external CVIs to evaluate internal CVIs reasonably. Finally, we validate the usefulness of the evaluation ability summary through four real-world datasets with various characteristics, indicating the effectiveness of CAMER.

This study relies on the experimental results to explore the evaluation abilities of external CVIs to internal CVIs in different cases. In the future, we will put effort into exploring their distinctions from a theoretical perspective.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 71971025 and China Scholarship Council under Grant 201906460085.

8 REFERENCES

[1] Oktar, Y. & Turkan, M. (2018). A review of sparsity-based clustering methods. *Signal Processing*, *148*(2018), 20-30. https://doi.org/10.1016/j.sigpro.2018.02.010

[2] Voican, O. (2020). Using Data Mining Methods To Solve Classification Problems In Financial-Banking Institutions. *Economic Computation & Economic Cybernetics Studies & Research*, *54*(1), 159-176. https://doi.org/10.24818/18423264/54.1.20.11

[3] Lei, Y., Bezdek, J. C., Romano, S., Vinh, N. X., Chan, J., & Bailey, J. (2017). Ground truth bias in external cluster validity indices. *Pattern Recognition*, *65*(2017), 58-70. https://doi.org/10.1016/j.patcog.2016.12.003

[4] Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Perez, J. M., & Perona, I. (2013). An extensive comparative study of cluster validity indices. *Pattern Recognition*, *46*(1), 243-256. https://doi.org/10.1016/j.patcog.2012.07.021

[5] Liu, Y., Li, Z., Xiong, H., Gao, X., Wu, J., & Wu, S. (2013). Understanding and enhancement of internal clustering validation measures. *IEEE transactions on cybernetics*, *43*(3), 982-994. https://doi.org/10.1109/TSMCB.2012.2220543

[6] Liu, Y., Li, Z., Xiong, H., Gao, X., & Wu, J. (2010). Understanding of internal clustering validation measures. *InProceedings of the 2010 IEEE International Conference on Data Mining*, 911-916. https://doi.org/10.1109/ICDM.2010.35

[7] Gurrutxaga, I., Muguerza, J., Arbelaitz, O., Pérez, J. M., & Martín, J. I. (2011). Towards a standard methodology to evaluate internal cluster validity indices. *Pattern Recognition Letters*, *32*(3), 505-515. https://doi.org/10.1016/j.patrec.2010.11.006

[8] Milligan, G. W. & Cooper, M. C. (1986). A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate behavioral research*, *21*(4), 441-458. https://doi.org/10.1207/s15327906mbr2104_5

[9] Romano, S., Bailey, J., Nguyen, V., & Verspoor, K. (2014, January). Standardized mutual information for clustering comparisons: one step further in adjustment for chance. *In Proceedings of theInternational Conference on Machine Learning*, 1143-1151.

[10] Vinh, N. X., Epps, J., & Bailey, J. (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *The Journal of Machine Learning Research*, *11*(2010), 2837-2854.

[11] Wu, J., Xiong, H., & Chen, J. (2009, June). Adapting the right measures for k-means clustering. *In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 877-886. https://doi.org/10.1145/1557019.1557115

[12] Wu, J., Yuan, H., Xiong, H., & Chen, G. (2010). Validation of overlapping clustering: A random clustering perspective. *Information Sciences*, *180*(22), 4353-4369. https://doi.org/10.1016/j.ins.2010.07.028
[13] Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, *30*(1/2), 81-93. https://doi.org/10.1093/biomet/30.1-2.81
[14] Spearman, C. (1906). Footrule for measuring correlation. *British Journal of Psychology*, *2*(1), 89-108. https://doi.org/10.1111/j.2044-8295.1906.tb00174.x
[15] Xie, J., Xiong, Z. Y., Dai, Q. Z., Wang, X. X., & Zhang, Y. F. (2020). A new internal index based on density core for clustering validation. *Information Sciences*, *506*(2020), 346-365. https://doi.org/10.1016/j.ins.2019.08.029
[16] Hu, L. & Zhong, C. (2019). An internal validity index based on density-involved distance. *IEEE Access*, *7*(2019), 40038-40051. https://doi.org/10.1109/ACCESS.2019.2906949
[17] Gao, X. & Wu, S. (2019). CUBOS: An Internal Cluster Validity Index for Categorical Data. *Tehnički vjesnik*, *26*(2), 486-494. https://doi.org/10.17559/TV-20190109015453
[18] Fu, L. & Wu, S. (2019). A new internal clustering validation index for categorical data based on concentration of attribute values. *Chinese Journal of Engineering*, *41*(5), 682-693.
[19] Cheng, D., Zhu, Q., Huang, J., Wu, Q., & Yang, L. (2018). A novel cluster validity index based on local cores. *IEEE transactions on neural networks and learning systems*, *30*(4), 985-999. https://doi.org/10.1109/TNNLS.2018.2853710
[20] Gao, X. & Yang, M. (2018). Understanding and enhancement of internal clustering validation indexes for categorical data. *Algorithms*, *11*(11), 177. https://doi.org/10.3390/a11110177
[21] Kim, B., Lee, H., & Kang, P. (2018). Integrating cluster validity indices based on data envelopment analysis. *Applied Soft Computing*, *64*(2018), 94-108. https://doi.org/10.1016/j.asoc.2017.11.052
[22] Zhou, S. & Xu, Z. (2018). A novel internal validity index based on the cluster centre and the nearest neighbour cluster. *Applied soft computing*, *71*(2018), 78-88. https://doi.org/10.1016/j.asoc.2018.06.033
[23] Zhao, X., Liang, J., & Dang, C. (2017). Clustering ensemble selection for categorical data based on internal validity indices. *Pattern Recognition*, *69*(2017), 150-168. https://doi.org/10.1016/j.patcog.2017.04.019
[24] Rojas-Thomas, J. C., Santos, M., & Mora, M. (2017). New internal index for clustering validation based on graphs. *Expert Systems with Applications*, *86*(2017), 334-349. https://doi.org/10.1016/j.eswa.2017.06.003
[25] Zhou, S., Xu, Z., & Liu, F. (2016). Method for determining the optimal number of clusters based on agglomerative hierarchical clustering. *IEEE Transactions on Neural Networks and learning systems*, *28*(12), 3007-3017. https://doi.org/10.1109/TNNLS.2016.2608001
[26] Fu, L. & Wu, S. (2016). An Internal Clustering Validation Index for Boolean Data. *Cybernetics and Information Technologies*, *16*(6), 232-244. https://doi.org/10.1515/cait-2016-0091
[27] Fränti, P. & Sieranoja, S. (2018). K-means properties on six clustering benchmark datasets. *Applied Intelligence*, *48*(12), 4743-4759. https://doi.org/10.1007/s10489-018-1238-7
[28] Dua, D. & Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. *Irvine, CA: University of California, School of Information and Computer Science*.
[29] Maaten, L. V. D. & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, *9*(2008), 2579-2605.

Contact information:

Xiaonan GAO, PhD candidate
School of Economics and Management,
University of Science and Technology Beijing,
30 Xueyuan Road, Haidian District, Beijing 100083, China
E-mail: gaoxiaonan0001@163.com

Guiying WEI, PhD,Associate Professor
School of Economics and Management,
University of Science and Technology Beijing,
30 Xueyuan Road, Haidian District, Beijing 100083, China
E-mail: weigy@manage.ustb.edu.cn

Sen WU, PhD, Full Professor
(Corresponding author)
School of Economics and Management,
University of Science and Technology Beijing,
30 Xueyuan Road, Haidian District, Beijing 100083, China
E-mail: wusen@manage.ustb.edu.cn

Falong FAN, undergraduate student
School of Economics and Management,
University of Science and Technology Beijing,
30 Xueyuan Road, Haidian District, Beijing 100083, China
E-mail: 13432921186@163.com