ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)

https://doi.org/10.17559/TV-20190109015453
Original scientific paper

CUBOS: An Internal Cluster Validity Index for Categorical Data

Xiaonan GAO, Sen WU

Abstract: Internal cluster validity index is a powerful tool for evaluating clustering performance. The study on internal cluster validity indices for categorical data has been a challenging task due to the difficulty in measuring distance between categorical attribute values. While some efforts have been made, they ignore the relationship between different categorical attribute values and the detailed distribution information between data objects. To solve these problems, we propose a novel index called Categorical data cluster Utility Based On Silhouette (*CUBOS*). Specifically, we first make clear the superiority of the paradigm of Silhouette index in exploring the details of clustering results. Then, we raise the Improved Distance metric for Categorical data (IDC) inspired by Category Distance to measure distance between categorical data exactly. Finally, the paradigm of Silhouette index and IDC are combined to construct the *CUBOS*, which can overcome the aforementioned shortcomings and produce more accurate evaluation results than other baselines, as shown by the experimental results on several UCI datasets.

Keywords: categorical data; clustering; distance metric; evaluation; internal cluster validity index

1 INTRODUCTION

Clustering is one of the most important tasks in data mining and machine learning that partitions dataset into different clusters in which data objects are similar to those in the same cluster and dissimilar to those in different clusters, to identify the nature structures and mine the potential useful information hidden under mass data [1]. It has been applied in many real-world domains, including pattern recognition [2], customer segmentation [3], anomaly detection [4, 5] and trending topic detection [6], et al. Since most of data in real-world lacks labels or other external information, it is hard to identify which clustering algorithms or parameter configurations yield the optimal clustering result. To this end, internal cluster validity indices, which evaluate the clustering performance without reference labels or other external information besides the structure of clustering results, have attracted lots of researchers' attentions [7].

Internal cluster validity indices are used to evaluate the clustering performance by considering only the clustering data, which can be briefly classified into numerical data-specific method and categorical data-specific method. Numerical data-specific method refers to the internal cluster validity indices that are applied to evaluate the clustering performance of numerical data. And categorical data-specific method refers to another kind of indices that are used to evaluate the clustering performance of categorical data.

The numerical data-specific method has been studied relatively adequately that evaluates clustering results according to the compactness of intra-cluster and separation of inter-clusters. Lots of internal cluster validity indices for numerical data have been proposed, such as Dunn index (*D*) [8], Calinski-Harabasz index (*CH*) [9], I index [10], Davies-Bouldin index (*DB*) [11] and Silhouette index (*S*) [12], et al. These indices measure the compactness of intra-cluster and separation of inter-clusters by computing the distance between numerical data objects or centroids, that are able to reflect the microscopic distribution information between data objects in clustering results and produce relatively more accurate evaluation results [13].

For categorical data, it is difficult to compute distance straightforward. The method used to measure the similarity or dissimilarity between two categorical data objects or of a categorical cluster can be divided into three types: simple matching-based approach, probability-based approach and entropy-based approach. Simple matching-based approach is to compute the dissimilarity between two categorical data objects according to whether the attribute values are identical, which is used in the well-known K-modes algorithms [14] typically. Probability-based approach is to measure the similarity or dissimilarity of a categorical cluster by computing the probability of identical attribute values of data objects in the cluster. In addition, entropy-based approach is relying on the association between entropy and cluster: there is a lower entropy in the cluster of similar data objects than in the cluster of dissimilar data objects. COOLCAT is a traditional entropy-based categorical data clustering algorithm [15]. The three types of measurement approaches are essentially rooted in the identity of categorical attribute values. Moreover, most of the existing internal cluster validity indices for categorical data rely on these similarity or dissimilarity measurement approaches.

There are some researches about internal cluster validity indices for categorical data, such as Cluster Cardinality Index (*CCI*) [16], Categorical Data Clustering with Subjective factors (*CDCS*) [17], Information Entropy (*IE*) [15], Category Utility (*CU*) [18] and New Condorcet Criterion (*NCC*) [19], et al. Among them, *CCI* and *NCC* rely on the simple matching-based approach to measure the compactness and separation, *CDCS* and *CU* rely on the probability-based approach and *IE* relies on the entropy-based approach. Since the kernel of these approaches is the identity of categorical attribute values, which leads to two deficiencies among the existing internal cluster validity indices for categorical data. One is that the indices only take into account the otherness among attribute values, not considering the relationship between different attribute values. The other is that most of the existing internal cluster validity indices for categorical data measure the compactness and separation only according to the similarity or dissimilarity of a cluster, cannot measure the similarity or dissimilarity between two categorical data objects, so that more detailed information of clustering results is unable to be explored. In this paper, we limit our scope to the improvement of internal cluster validity indices for categorical data to overcome the two deficiencies.

For exploring more details hidden in clustering results of categorical data, we develop a distance metric for categorical data, called Improved Distance metric for Categorical data (*IDC*), which can compute distance between two categorical data objects considering the relationship between different attribute values. Moreover, the paradigm of Silhouette (*S*) index which obtains significantly better evaluation results than other existing internal cluster validity indices for numerical data [20] is used to construct a novel internal cluster validity index for categorical data, called Categorical data cluster Utility Based On Silhouette (*CUBOS*), with the proposed *IDC*.

The main contributions of this paper are summarized as follows. Above all, we analyse the characteristics of several existing representative internal cluster validity indices for numerical data and illustrate the essence of each index in a visual way to demonstrate the superiority of Silhouette (*S*) index. In addition, we develop a novel distance metric for categorical data *IDC* under the inspiration of Category Distance that has been presented in an existing work [21], which satisfies the distance conditions (non-negativity, symmetry and triangular inequality). The proposed distance metric *IDC* computes the distance between two categorical data objects considering the relationship between different attribute values. Finally, an internal cluster validity index for categorical data *CUBOS* is proposed which combines the *IDC* and the paradigm of *S* index, not only realizes the accurate measurement of the distance between two categorical data objects, but also explores detailed distribution information in clustering results.

2 RELATED WORK

In this section, we review several typical internal cluster validity indices for numerical data and categorical data and analyse their respective characteristics.

2.1 Internal Cluster Validity Indices for Numerical Data

Let $X=\{x_1,x_2,\cdots,x_i,\cdots,x_n\}$ be a numerical dataset of *n* data objects with *m* attributes. $\pi=\{C_1,C_2,\cdots,C_k\}$ is a clustering result of dataset *X*, where *k* is the number of clusters. The number of data objects in cluster $C_j, j\in\{1,2,\cdots,k\}$ is $|C_j|$. The data objects in cluster C_j are $C_j=\left\{x_{c_j}^1,x_{c_j}^2,\cdots,x_{c_j}^{|c_j|}\right\}$. *c* is the centroid of dataset *X*, c_j is the centroid of cluster C_j.

(1) Dunn index (*D*)

Dunn index is formulated as follows:

$$D(\pi)=\min_{1\le j\le k}\left\{\min_{1\le r\le k,r\ne j}\left\{\frac{\min\limits_{1\le g\le|C_j|,1\le h\le|C_r|} d\left(x_{C_j}^g,x_{C_r}^h\right)}{\max\limits_{1\le s\le k}\left\{\max\limits_{1\le p\le|C_s|,1\le q\le|C_s|} d\left(x_{C_s}^p,x_{C_s}^q\right)\right\}}\right\}\right\} \quad (1)$$

where the numerator represents the separation of inter-clusters by computing the minimum distance between two data objects in different clusters and the denominator represents the compactness of intra-cluster by computing the maximum distance between two data objects in the same cluster. It is easy to see that large *D* value indicates good clustering performance.

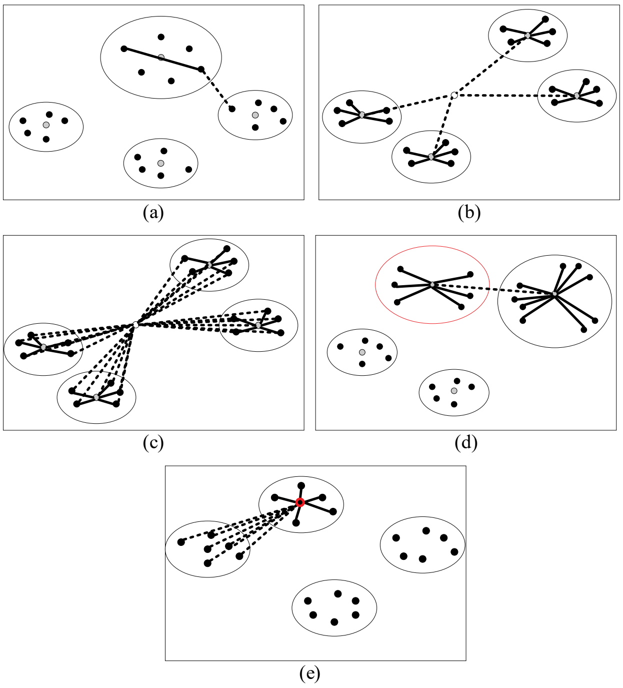

Figure 1 Distribution diagrams

The distribution diagram of *D* index is shown in Fig. 1 (a). There are four clusters, the black points in each circle represent the data objects belonging to that cluster and the gray point represents the centroid of each cluster. The solid straight line and the dotted straight line are respectively used to indicate the compactness of intra-cluster and the separation of inter-clusters. It is obvious that *D* index evaluates the clustering performance based only on two distances, namely the maximum distance in a cluster and the minimum distance between clusters, without considering other distribution information, which results in the relatively inaccurate evaluation results.

(2) Calinski-Harabasz index (*CH*)

Calinski-Harabasz index is given as follows:

$$CH(\pi)=\frac{\frac{1}{k-1}\sum\limits_{i=1}^{k}|C_i|d^2(c_i,c)}{\frac{1}{n-k}\sum\limits_{j=1}^{k}\sum\limits_{g=1}^{|C_j|}d^2\left(x_{C_j}^g,c_j\right)} \quad (2)$$

where the numerator represents the separation of inter-clusters by computing the weighted average of the square of distance from the centroid of each cluster to the centroid of dataset, and the denominator represents the compactness of intra-cluster by computing the square of distance from each data object in a cluster to its centroid. Similarly, large *CH* value indicates good clustering performance.

The distribution diagram of *CH* index is shown in Fig. 1(b). The white point is the centroid of dataset. Compared with *D* index, *CH* index considers the distribution of all data objects. However, *CH* index only focuses on the centroid-based relationship, e.g. the distance from data object to its centroid and the distance from the centroid of cluster to the centroid of dataset, but not on the relationship between data objects, which leads to that *CH* index cannot accurately evaluate the clustering performance in some cases. For example, on one side, *CH* index might misjudge that the separation of inter-clusters is good where each cluster is far from the centroid of dataset but some clusters are close to each other, on the other side, the compactness of intra-cluster may be misjudged as good when each data objects is close to its corresponding centroid but some of them are far away.

(3) I index (*I*)

I index is defined as follows:

$$I(\pi)=\left[\frac{1}{k}\frac{\sum_{i=1}^{n}d\left(x_i,c\right)}{\sum_{j=1}^{k}\sum_{g=1}^{\left|C_j\right|}d\left(x_{C_j}^{g},c_j\right)}\max_{\substack{1\le s\le k,\\1\le r\le k,\\s\ne r}}d\left(c_s,c_r\right)\right]^{p} \tag{3}$$

where, the separation of inter-clusters is measured according to the distance from each data object to the centroid of dataset and the maximum distance between centroids of clusters, and the compactness of intra-cluster is measured by computing the distance between data object and its corresponding centroid. The maximum *I* index value indicates the optimal clustering result.

The distribution diagram of *I* index is shown in Fig. 1(c). It is very similar to the distribution diagram of *CH* index, except that *I* index also measures the distance from each data object to the centroid of dataset. Although *I* index considers more distribution information than *CH* index, it still evaluates the clustering performance based on the centroid-based distance like *CH* index, this kind of distance metric results in the neglect of the relationship between clusters or data objects.

(4) Davies-Bouldin index (*DB*)

Davies-Bouldin index is given as follows:

$$DB(\pi)=\frac{1}{k}\sum_{s=1}^{k}\max_{1\le r\le k,r\ne s}\left(\frac{\frac{1}{\left|c_s\right|}\sum_{g=1}^{\left|c_s\right|}d\left(x_{c_s}^{g},c_s\right)+\frac{1}{\left|c_r\right|}\sum_{h=1}^{\left|c_r\right|}d\left(x_{c_r}^{h},c_r\right)}{d\left(c_s,c_r\right)}\right) \tag{4}$$

where the *DB* index evaluates the clustering performance by measuring the performance of each cluster respectively based on the average similarity of the cluster with its most similar cluster. Small *DB* index value indicates good clustering performance.

The distribution diagram of *DB* index is shown in Fig. 1(d). It only shows the distance between data objects or centroids involved in measuring the performance of one cluster. *DB* index measures the compactness of intra-cluster by computing the distance from data objects to the centroid in the same way as *CH* index. Similarly, this kind of method would produce an inaccurate evaluation result. Furthermore, it measures the separation of inter-clusters by computing the distance between centroids, which cannot evaluate the distance between two clusters exactly, for example, when the centroids of the two clusters are far away but their boundaries are actually close to each other as shown in Fig. 1(d).

(5) Silhouette index (*S*)

Silhouette index is formulated as follows:

$$S(\pi)=\frac{1}{k}\sum_{i=1}^{k}\left(\frac{1}{\left|C_i\right|}\sum_{g=1}^{\left|C_i\right|}\frac{b\left(x_{C_i}^{g}\right)-a\left(x_{C_i}^{g}\right)}{\max\left[b\left(x_{C_i}^{g}\right),a\left(x_{C_i}^{g}\right)\right]}\right) \tag{5}$$

where

$$a\left(x_{C_i}^{g}\right)=\frac{1}{\left|C_i\right|-1}\sum_{\substack{h=1\\h\ne g}}^{\left|C_i\right|}d\left(x_{C_i}^{g},x_{C_i}^{h}\right),$$

$$b\left(x_{C_i}^{g}\right)=\min_{1\le j\le k,j\ne i}\left(\frac{1}{\left|C_j\right|}\sum_{h=1}^{\left|C_j\right|}d\left(x_{C_i}^{g},x_{C_j}^{h}\right)\right).$$

S index evaluates the clustering performance by measuring the performance of each data object. The compactness of a data object is computed by the average distance from the data object to other data objects in the same cluster. In addition, the separation of a data object is the minimum of average distance from the data object to data objects in another cluster. Large *S* index value indicates good clustering performance.

The distribution diagram of *S* index is shown in Fig. 1(e). It only shows the distance related to the compactness and separation of a data object which is a black point with red edge. We can see that the distances computed in *S* index are between data object, but not related to the centroids, and compared to *D* index, more data objects are taken into account. Therefore, *S* index considers more distribution information and can produce much more accurate evaluation results.

After introducing the above five typical internal cluster validity indices for numerical data, we can know that *CH*, *I* and *DB* indices evaluate clustering performance through the centroid-based distance, *D* and *S* indices evaluate clustering performance through the data object-based distance. Since the centroid-based distance neglects the relationship between data objects resulting in the inaccurate reflection for the true distribution of clustering results, *CH*, *I* and *DB* indices cannot produce precise evaluation results for clustering results. Although *D* index is based on the data object-based distance, it only takes into account a little distribution information of clustering results, which leads to defective reflection for the overall distribution of clustering results. *S* index evaluates the clustering performance of each data object based on the data object-based distance and the true distribution information can be reflected as much as possible, hence *S* index can produce much more precise evaluation results than other indices. Based on this, we exploit the paradigm of *S* index to construct a novel internal cluster validity index for categorical data.

2.2 Internal Cluster Validity Indices for Categorical Data

Let $X=\left\{x_1,x_2,\cdots,x_i,\cdots,x_n\right\}$ be a categorical dataset of *n* data objects with *m* attributes $A=\left[a_1,a_2,\cdots,a_m\right]$.

$V_X^d = \left\{ v_{Xd}^1, v_{Xd}^2, \cdots, v_{Xd}^{\left|V_X^d\right|} \right\}$ is the set of values on attribute $a_d, 1 \le d \le m$ for categorical dataset *X*, $\left|V_X^d\right|$ is the number of values. $\pi = \{C_1, C_2, \cdots, C_k\}$ is a clustering result of dataset *X*, where *k* is the number of clusters. $\left|C_i\right|$ is the number of data objects in cluster $C_i, i \in \{1, 2, \cdots, k\}$.

(1) Cluster Cardinality Index (*CCI*)

Cluster Cardinality index is formulated as follows:

$$CCI(\pi) = \frac{1}{k}\sum_{i=1}^{k} \max_{1 \le j \le k, j \ne i} \left\{ \frac{CI(i) + CI(j)}{CI(i,j)} \right\} \tag{6}$$

where

$$CI(i) = \frac{1}{m}\sum_{d=1}^{m} \frac{\left|V_{C_i}^d\right|}{\left|C_i\right|},$$

$$CI(i,j) = \frac{1}{m}\sum_{d=1}^{m} \frac{\left|V_{C_i}^d \cup V_{C_j}^d\right| - \left|V_{C_i}^d \cap V_{C_j}^d\right| + 1}{\left|V_{C_i}^d \cup V_{C_j}^d\right| + 1}.$$

CI(*i*) and *CI*(*i*, *j*) respectively denote the dissimilarity of data objects in cluster C_i and the dissimilarity between cluster C_i and C_j. Apparently, small *CCI* index value indicates good clustering performance.

(2) Categorical Data Clustering with Subjective factors (*CDCS*)

Categorical Data Clustering with Subjective factors is defined as follows:

$$CDCS(\pi) = \frac{intra(\pi)}{inter(\pi)} \tag{7}$$

where *intra*(*π*) represents the compactness of intra-cluster for clustering results, which is computed as follows:

$$intra(\pi) = \sum_{i=1}^{k} \frac{\left|C_i\right|}{n} \sum_{d=1}^{m} \frac{1}{m} \left(\max_{q}^{\left|V_X^d\right|} P\left(a_d = v_{C_i d}^q\right) \right)^3 \tag{8}$$

where $P\left(a_d = v_{C_i d}^q\right)$ denotes the probability of $v_{C_i d}^q$ on attribute a_d in cluster C_i. Moreover, *inter*(*π*) represents the separation of inter-clusters for clustering results, which is defined as follows:

$$inter(\pi) = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} \left(Sim\left(C_i, C_j\right)^{\frac{1}{m}} \left|C_i \cup C_j\right| \right)}{(k-1) \times n} \tag{9}$$

where $Sim\left(C_i, C_j\right)$ is the similarity between cluster C_i and cluster C_j that is computed as follows:

$$Sim\left(C_i, C_j\right) = \prod_{d=1}^{m} \left[\sum_{q=1}^{\left|V_X^d\right|} \min\left\{ P\left(a_d = v_{Xd}^q \middle| C_i\right), P\left(a_d = v_{Xd}^q \middle| C_j\right) \right\} + \varepsilon \right] \tag{10}$$

$P\left(a_d = v_{Xd}^q \middle| C_i\right)$ denotes the probability of v_{Xd}^q on attribute a_d in cluster C_i. The idea of measuring the similarity between two categorical clusters is that the more the identical attribute values of the two clusters, the more similar they are. According to the equations and their description, the best clustering results would be indicated by the largest *CDCS* values.

(3) Information Entropy (*IE*)

Information Entropy is given as follows:

$$IE(\pi) = \frac{1}{k}\sum_{i=1}^{k} \frac{\left|C_i\right|}{n} \sum_{d=1}^{m} \sum_{q=1}^{\left|V_X^d\right|} P\left(a_d = v_{Xd}^q \middle| C_i\right) \log P\left(a_d = v_{Xd}^q \middle| C_i\right) \tag{11}$$

IE index evaluates the clustering performance by exploiting information entropy theory. The basic idea is that the entropy of cluster within similar data objects is lower than that of cluster within dissimilar data objects. It is obvious that smaller *IE* index values indicate better clustering results.

(4) Category Utility (*CU*)

Category Utility is defined as follows:

$$CU(\pi) = \frac{1}{k}\sum_{i=1}^{k} \frac{\left|C_i\right|}{n} \sum_{d=1}^{m} \sum_{q=1}^{\left|V_X^d\right|} \left[P\left(a_d = v_{Xd}^q \middle| C_i\right)^2 - P\left(a_d = v_{Xd}^q\right)^2 \right] \tag{12}$$

CU index tries to evaluate the clustering performance by measuring the identity of attribute values in a cluster. Larger value of *CU* index indicates better clustering result.

(5) New Condorcet Criterion (*NCC*)

New Condorcet Criterion is formulated as follows:

$$NCC(\pi) = \sum_{i=1}^{k} \left(S_{intra}\left(C_i\right) + D_{inter}\left(C_i\right) \right) \tag{13}$$

where $S_{intra}(C_i)$ denotes the compactness of intra-cluster for cluster C_i, which is computed as follows:

$$S_{intra}\left(C_i\right) = \sum_{x_{C_i}^j \in C_i} \sum_{\substack{x_{C_i}^g \in C_i \\ g \ne j}} \left(m - d_{jg}\right) \tag{14}$$

where $d_{C_i}^{jg}$ is the number attributes with different values for data objects $x_{C_i}^j$ and $x_{C_i}^g$ in cluster C_i. And $D_{inter}(C_i)$ denotes the separation of inter-clusters for the cluster C_i, which is computed as follows:

$$D_{inter}\left(C_i\right) = \sum_{x_{C_i}^j \in C_i} \sum_{x_{C_i}^g \notin C_i} \left(d_{jg}\right) \tag{15}$$

Apparently, larger *NCC* index values indicate better clustering results.

From the above description about internal cluster validity indices for categorical data, we can know two facts. One is that *CCI*, *CDCS*, *IE* and *CU* indices measure similarity or dissimilarity based on probability according to their definitions. This similarity or dissimilarity measurement method only pays attention to the number of occurrences of attribute values, ignoring the relationship between different attribute values. *NCC* index measures the distance between two data objects based on the matching of all attribute values compared, which is also incapable to identify the relationship between different attribute values. The other is that *CCI*, *CDCS*, *IE* and *CU* indices evaluate the clustering performance based on the similarity or dissimilarity of a cluster, but do not measure the similarity or dissimilarity of data objects more meticulously. Therefore, more detailed distribution information of clustering results cannot be discovered.

After analyzing the characteristics of several typical existing internal cluster validity indices for numerical and categorical data, we realize that Silhouette (*S*) index can obtain more accurate evaluation results for clustering performance compared with other internal cluster validity indices for numerical data. Therefore, we exploit the paradigm of *S* index to construct a new internal cluster validity index for categorical data. In addition, there are two deficiencies of most of the existing internal cluster validity indices for categorical data. One is the overlook of relationship between different attribute values. The other is the incapability of discovering more detailed distribution information between data objects. To overcome the two deficiencies, a new distance metric for categorical data IDC is proposed that can reflect the relationship between different attribute values and satisfy the distance conditions. By using this distance metric, we can explore more detailed distribution information in the clustering results. Moreover, a novel internal cluster validity index for categorical data *CUBOS* is developed by combining the IDC and the paradigm of *S* index.

3 CATEGORICAL DATA CLUSTER UTILITY BASED ON SILHOUETTE

Our proposed internal cluster validity index for categorical data *CUBOS* consists of two components: (a) presenting the Improved Distance metric for Categorical data (*IDC*) inspired by the Category Distance in an existing related work; (b) constructing the new internal cluster validity index *CUBOS* by combining the presented *IDC* and the paradigm of Silhouette index. Specifically, to illustrate our presented index clearly, the Category Distance that inspires our research will be reviewed and discussed firstly.

3.1 Discussion on Category Distance

The Category Distance has been proposed in [21], which relies on the weights of values on each categorical attribute, and no longer depends on the independence assumption that there is no relationship between the values on the same attribute. To define a distance formula satisfying the distance conditions that consist of non-negativity, symmetry and triangular inequality, a general distance metric paradigm has been provided as follows:

$$\psi\left(c,c'\right)=\begin{cases}\sqrt{\rho\left(c\right)}, c=c' \\ \sqrt{\frac{1}{2}\left[\bar{\rho}\left(c\right)+\bar{\rho}\left(c'\right)\right]}, c\neq c'\end{cases} \quad (16)$$

$$s.t.\forall c:\bar{\rho}\left(c\right)\geq\rho\left(c\right)\geq 0$$

where c and c' are two values on the same attribute, $\rho(c)$ denotes the dissimilarity of attribute value c when two data objects take identical value on the same attribute, correspondingly, $\bar{\rho}\left(c\right)$ denotes the dissimilarity of attribute value c when two data objects take different values covering c on the same attribute. It was proved that any distance metric meeting this paradigm satisfies the distance conditions.

Category Distance meeting the paradigm was proposed in their work, where $\rho\left(v_{Xd}^{l}\right)$ and $\bar{\rho}\left(v_{Xd}^{l}\right)$ for $\forall v_{Xd}^{l}\in V_{X}^{d}$ are formulated as follows:

$$\begin{cases}\rho\left(v_{Xd}^{l}\right)=\left[1-\lambda_{X}\left(v_{Xd}^{l}\right)\right]^{\frac{1}{\beta}} \\ \bar{\rho}\left(v_{Xd}^{l}\right)=1+\left[\lambda_{X}\left(v_{Xd}^{l}\right)\right]^{\frac{1}{\beta}}\end{cases}, \ l=1,2,\cdots\left|V_{X}^{d}\right| \quad (17)$$

Where $0\leq\lambda_{X}\left(v_{Xd}^{l}\right)\leq 1$ is the weight of attribute value v_{Xd}^{l} that reflects the contribution of attribute value v_{Xd}^{l} for the distance computation. The exponent $1/\beta > 1$ is used to control the strength of the contribution of attribute values. According to Eq. (16) and Eq. (17), the Category Distance was developed as follows:

$$\psi_{CD}\left(c,c'\right)=\begin{cases}\left[1-\lambda_{X}\left(c\right)\right]^{\frac{1}{2\beta}}, \ c=c' \\ \left[1+\frac{1}{2}\lambda_{X}\left(c\right)^{\frac{1}{\beta}}+\frac{1}{2}\lambda_{X}\left(c'\right)^{\frac{1}{\beta}}\right]^{\frac{1}{2}}, \ c\neq c'\end{cases} \quad (18)$$

In their work, the computation of distance metric in Eq. (18) was converted into an optimizing problem, the weights optimization of attribute values, which was solved by a clustering algorithm. That means the distance between two categorical attribute values cannot be computed directly but be fused in clustering procedure.

The Category Distance $\psi_{CD}\left(c,c'\right)$ discards the independence assumption between categorical attribute values to reflect the relationship between different values on the same categorical attribute and satisfies the three distance conditions, which can be applied flexibly into the paradigm of internal cluster validity indices for numerical data studied more fully. Nevertheless, there are a few defects of Category Distance. On one hand, the heterogeneity of data objects sharing the identical categorical attribute value exists according to Eq. (18), that causes the distance between two data objects with identical values on all categorical attributes to be greater than 0. On the other hand, the distance computation and the clustering procedure are integrated together, so it is impossible to compute distance separately for other tasks.

In this paper, we improve the Category Distance to overcome its two defects: firstly, we develop the computation method for weights of attribute values based on the whole dataset X under the assumption that uncommon attribute values contribute more weights than common attribute values. This idea is consistent with the information theory that the events with lower occurrence probability can provide more information than events with higher occurrence probability. Secondly, we adjust the general distance metric paradigm listed in Eq. (16) for that the distance between two data objects with identical values on all categorical attributes is 0. Finally, we propose the Improved Distance metric for Categorical data called *IDC* based on the developed computation method for weights of attribute values and the adjusted general distance metric paradigm.

3.2 Improved Distance Metric for Categorical Data

(1) Developed computation method for weights of attribute values

The developed computation method for weights of attribute values is shown as follows:

$$\lambda_X\left(v_{Xd}^l\right)=\sum_{v_{Xd}^q\in MSFVS\left(v_{Xd}^l\right)} p\left(v_{Xd}^q\right) \tag{19}$$

$$p\left(v_{Xd}^q\right)=\frac{f\left(v_{Xd}^q\right)\times\left(f\left(v_{Xd}^q\right)-1\right)}{n\times\left(n-1\right)} \tag{20}$$

where $f\left(v_{Xd}^q\right)$ is the occurrence number of value v_{Xd}^q in dataset X. $MSFVS\left(v_{Xd}^l\right)$ is the set of all values on d^{th} categorical attribute in dataset X whose probabilities are equivalent to or smaller than the probability of v_{Xd}^l.

The weight computation method is derived from the similarity measure proposed by Goodall [22] that reflects the relationship between different attribute values by giving grater weights to uncommon attribute values.

(2) Adjusted general distance metric paradigm

The adjusted paradigm is shown as follows:

$$\psi\left(c,c'\right)=\begin{cases}0, c=c'\\ \sqrt{\frac{1}{2}\left[\bar{\rho}\left(c\right)+\bar{\rho}\left(c'\right)\right]}, c\neq c'\end{cases} \tag{21}$$
$$s.t.\forall c:\bar{\rho}\left(c\right)\geq 0$$

We only change the distance between two identical attribute values to 0. Along this line, the distance between two data objects with identical values on all categorical attributes would equal to 0. Moreover, any distance metric applying this paradigm satisfies the distance conditions:

$$\begin{cases}\psi\left(a,b\right)\geq 0;\left(non-negativity\right)\\ \psi\left(a,b\right)=\psi\left(b,a\right);\left(symmetry\right)\\ \psi\left(a,b\right)\leq\psi\left(a,c\right)+\psi\left(c,b\right);\left(triangular\quad inequality\right)\end{cases} \tag{22}$$

where $\psi(a, b)$ is the distance between a and b. The Eq. (21) obviously follows the conditions of non-negativity and symmetry. For triangular inequality, we illustrate through five cases:

Case 1: $a = b = c$; According to Eq. (21), when $a = b$, there is $\psi(a, b) = 0$. Similarly, $\psi(a, c) = 0$ and $\psi(c, b) = 0$. Hence, $\psi(a, b) \leq \psi(a, c) + \psi(c, b)$;

Case 2: $a = b$, $a \neq c$ and $b \neq c$; We have $\psi(a, b) = 0$, $\psi\left(a,c\right)=\sqrt{\frac{1}{2}\left[\bar{\rho}\left(a\right)+\bar{\rho}\left(c\right)\right]}$ and $\psi\left(c,b\right)=\sqrt{\frac{1}{2}\left[\bar{\rho}\left(c\right)+\bar{\rho}\left(b\right)\right]}$. Since $0\leq\sqrt{\frac{1}{2}\left[\bar{\rho}\left(a\right)+\bar{\rho}\left(c\right)\right]}+\sqrt{\frac{1}{2}\left[\bar{\rho}\left(c\right)+\bar{\rho}\left(b\right)\right]}$, $\psi(a, b) \leq \psi(a, c) + \psi(c, b)$.

Case 3: $a = c$, $a \neq b$ and $b \neq c$; There are $\psi(a, c) = 0$, $\psi\left(a,b\right)=\sqrt{\frac{1}{2}\left[\bar{\rho}\left(a\right)+\bar{\rho}\left(b\right)\right]}$ and $\psi\left(c,b\right)=\sqrt{\frac{1}{2}\left[\bar{\rho}\left(c\right)+\bar{\rho}\left(b\right)\right]}$. Since $a = c$, $\sqrt{\frac{1}{2}\left[\bar{\rho}\left(a\right)+\bar{\rho}\left(b\right)\right]}=\sqrt{\frac{1}{2}\left[\bar{\rho}\left(c\right)+\bar{\rho}\left(b\right)\right]}$, thus, $\sqrt{\frac{1}{2}\left[\bar{\rho}\left(a\right)+\bar{\rho}\left(b\right)\right]}=0+\sqrt{\frac{1}{2}\left[\bar{\rho}\left(c\right)+\bar{\rho}\left(b\right)\right]}$, we have $\psi(a, b) \leq \psi(a, c) + \psi(c, b)$.

Case 4: $b = c$, $a \neq b$ and $b \neq c$; There are $\psi(c, b) = 0$, $\psi\left(a,b\right)=\sqrt{\frac{1}{2}\left[\bar{\rho}\left(a\right)+\bar{\rho}\left(b\right)\right]}$ and $\psi\left(a,c\right)=\sqrt{\frac{1}{2}\left[\bar{\rho}\left(a\right)+\bar{\rho}\left(c\right)\right]}$. Since $b = c$, $\sqrt{\frac{1}{2}\left[\bar{\rho}\left(a\right)+\bar{\rho}\left(b\right)\right]}=\sqrt{\frac{1}{2}\left[\bar{\rho}\left(a\right)+\bar{\rho}\left(c\right)\right]}$, thus, $\sqrt{\frac{1}{2}\left[\bar{\rho}\left(a\right)+\bar{\rho}\left(b\right)\right]}=\sqrt{\frac{1}{2}\left[\bar{\rho}\left(a\right)+\bar{\rho}\left(c\right)\right]}+0$, we have $\psi(a, b) \leq \psi(a, c) + \psi(c, b)$.

Case 5: $a \neq b$, $a \neq c$ and $b \neq c$; There are $\psi\left(a,b\right)=\sqrt{\frac{1}{2}\left[\bar{\rho}\left(a\right)+\bar{\rho}\left(b\right)\right]}$, $\psi\left(a,c\right)=\sqrt{\frac{1}{2}\left[\bar{\rho}\left(a\right)+\bar{\rho}\left(c\right)\right]}$ and $\psi\left(c,b\right)=\sqrt{\frac{1}{2}\left[\bar{\rho}\left(c\right)+\bar{\rho}\left(b\right)\right]}$. Since $\psi\left(a,b\right)^2=\frac{1}{2}\left[\bar{\rho}\left(a\right)+\bar{\rho}\left(b\right)\right]$ and $\left[\psi\left(a,c\right)+\psi\left(c,b\right)\right]^2=\frac{1}{2}\left[\bar{\rho}\left(a\right)+\bar{\rho}\left(c\right)\right]+\frac{1}{2}\left[\bar{\rho}\left(c\right)+\bar{\rho}\left(b\right)\right]+$ $+\frac{1}{2}\sqrt{\frac{1}{2}\left[\bar{\rho}\left(a\right)+\bar{\rho}\left(c\right)\right]}\times\sqrt{\frac{1}{2}\left[\bar{\rho}\left(c\right)+\bar{\rho}\left(b\right)\right]}$, hence, $\psi(a, b)^2 \leq [\psi(a, c) + \psi(c, b)]^2$, we have $\psi(a, b) \leq \psi(a, c) + \psi(c, b)$.

We can raise a distance metric for categorical data based on the developed computation method for weights of attribute values and the adjusted general distance metric paradigm to be applied in the S index for evaluating clustering performance.

(3) Improved Distance metric for Categorical data (*IDC*)

The Improved Distance metric for Categorical data (*IDC*) is raised as follows:

$$IDC\left(x_i,x_j\right)=\sum_{d=1}^{m}\psi_{IDC}\left(x_i^d,x_j^d\right) \tag{23}$$

$$\psi_{IDC}\left(c,c'\right)=\begin{cases} 0, c=c' \\ \sqrt{1+\frac{1}{2}\lambda_X\left(c\right)^{\frac{1}{\beta}}+\frac{1}{2}\lambda_X\left(c'\right)^{\frac{1}{\beta}}}, c\neq c' \end{cases} \tag{24}$$

where $\lambda_X(c)$ can be computed according to Eq. (19) and Eq. (20). The $1/\beta$ is used to control the strength of weights.

IDC discards the assumption that the different values on the same attribute are independent of each other and can express their relationship. Additionally, *IDC* satisfies the distance conditions which can be applied directly in the existing internal cluster validity indices based on distance.

3.3 Categorical Data Cluster Utility Based on Silhouette

Considering the superiority of Silhouette (*S*) index over other internal cluster validity indices for numerical data, we combine the *IDC* and the paradigm of *S* index to develop an internal cluster validity index for categorical data named Categorical data cluster Utility Based On Silhouette (*CUBOS*), that is defined as follows:

$$CUBOS\left(\pi\right)=\frac{1}{k}\sum_{i=1}^{k}\left(\frac{1}{\left|C_i\right|}\sum_{g=1}^{\left|C_i\right|}\frac{b\left(x_{C_i}^g\right)-a\left(x_{C_i}^g\right)}{\max\left[b\left(x_{C_i}^g\right),a\left(x_{C_i}^g\right)\right]}\right) \tag{25}$$

$$a\left(x_{C_i}^g\right)=\frac{1}{\left|C_i\right|-1}\sum_{\substack{h=1\\h\neq g}}^{\left|C_i\right|}IDC\left(x_{C_i}^g,x_{C_i}^h\right) \tag{26}$$

$$b\left(x_{C_i}^g\right)=\min_{1\leq j\leq k,j\neq i}\left(\frac{1}{\left|C_j\right|}\sum_{h=1}^{\left|C_j\right|}IDC\left(x_{C_i}^g,x_{C_j}^h\right)\right) \tag{27}$$

CUBOS index inherits the strength of *S* index that evaluates the clustering performance depending on the data object-based distance to expose as much as possible the more detailed distribution information in clustering results. Besides, *IDC* used in *CUBOS* index can compute the exact distance between two categorical data objects satisfying the distance conditions, rather than just estimate their similarity or dissimilarity. Meanwhile, *IDC* considers the relationship between different values on the same categorical attribute no longer based on the independence assumption.

4 EXPERIMENTAL RESULTS

Extensive experiments on several datasets from *UCI* are conducted to illustrate the effectiveness of *CUBOS*.

4.1 Experimental Datasets

Five typical categorical datasets from *UCI* are selected in the experiments. Tab. 1 lists these datasets.

Table 1 Summary of datasets

Abbr	Dataset Name	#Instances	#Attributes	#Clusters
BC	Breast Cancer	286	9	2
D	Dermatology	366	35	6
MB	Molecular Biology	106	57	2
S	Soybean	47	35	4
CVR	Congressional Voting Records	435	16	2

Specifically, there are missing values in *BC* dataset and *CVR* dataset. We delete the data objects containing missing values before clustering.

4.2 Evaluation Metrics

External cluster validity indices are to assess the consistency between clustering labels and true labels that can be used to evaluate the performance of internal indices. However, different external indices would lead to different measurement results for the same clustering results. Therefore, we exploit seven external cluster validity indices, including Accuracy (*A*), Adjusted Rand Index (*ARI*), F-measure (*F*), Micro-p (*M*), Normalized Mutual Information (*NMI*), Purity (*P*) and Rand Index (*RI*) [23, 24], to evaluate the performance of clustering results selected by internal indices, as shown in Tab. 2.

Table 2 External cluster validity indices used in the experiments

Abbr	Description	Formula	Direction
A	Accuracy	$\frac{1}{n}\sum_{j=1}^{kt}n_{jj}$	↑
ARI	Adjusted rand index	$\frac{RI-E\left[RI\right]}{\max\left(RI\right)-E\left[RI\right]}$	↑
F	F-measure	$\sum_{i=1}^{kc}\frac{n_i}{n}\times\max_{1\leq j\leq kt}\left\{\frac{2\cdot\frac{n_{ij}}{n_i}\cdot\frac{n_{ij}}{n_j}}{\frac{n_{ij}}{n_i}+\frac{n_{ij}}{n_j}}\right\}$	↑
M	Micro-p	$\frac{\sum_{j=1}^{kt}\max_{1\leq i\leq kc}n_{ij}}{n}$	↑
NMI	Normalized mutual information	$\frac{\sum_{1\leq i\leq kc}\sum_{1\leq j\leq kt}n_{ij}\log_2\frac{n_{ij}}{n_i n_j}}{\sqrt{\sum_{1\leq i\leq kc}n_i\log_2\frac{n_i}{n}}\sqrt{\sum_{1\leq j\leq kt}n_j\log_2\frac{n_j}{n}}}$	↑
P	Purity	$\frac{1}{n}\sum_{i=1}^{kc}\max_{1\leq j\leq kt}\left\|C_i\cap CT_j\right\|$	↑
RI	Rand index	$\frac{a+b}{n(n-1)/2}$	↑

In Tab. 2, a dataset with *n* data objects from *kt* classes $\theta=\left\{CT_1,CT_2,\cdots,CT_{kt}\right\}$ is partitioned into *kc* clusters $\pi=\left\{C_1,C_2,\cdots,C_{kc}\right\}$. The number of data objects which are from class *j* and partitioned into cluster *i* is n_{ij}. Additionally, *a* refers to the number of data object pairs that belong to different classes and are still clustered into different clusters. *b* refers to the number of data object pairs that belong to the same class and are still clustered into the same cluster. In addition, the larger the values of external indices are, the better the performance of clustering results chosen by internal indices for categorical data.

4.3 Baselines and Experimental Configurations

We compare the proposed *CUBOS* index with five baselines, which are introduced as Tab. 3:

K-modes algorithm is used to conduct clustering with the number of clusters ranging from 2 to $\sqrt{n}$, where *n* is the number of data objects in the dataset. And we preset the parameter β ={0.05, 0.1, 0.15,…, 0.95} for *CUBOS* index.

Table 3 Summary of algorithms to be compared

Index name	Source	Role	Direction
CUBOS	Section 3.3	Index proposed	↑
CCI	Gao et al. [16]	Baseline	↓
CDCS	Chang et al. [17]	Baseline	↑
IE	Barbara et al. [15]	Baseline	↓
CU	Gluck. [18]	Baseline	↑
NCC	Pierre. [19]	Baseline	↑

4.4 Performance Comparison

The evaluation results are reported in Tab. 4 to Tab. 10. The decimals in the tables are the evaluation scores for the performance of clustering results chosen by each internal index, and the integers in brackets indicate the ranking of effectiveness of internal indices.

First of all, we focus on the evaluation results with *A* as metric. In Tab. 4, we can see that *CUBOS*, *IE* and *CU* obtain better evaluation results than other indices and the performance of *CCI* is the worst. Although there are three times for being ranking first of *CUBOS*, *IE* and *CU*, *CUBOS*'s rankings on the remaining two datasets are respectively second and third which are in front of the rankings of *IE* and *CU* on their remaining two datasets. Thus, the effectiveness of *CUBOS* is relatively superior than that of other indices with *A* as metric.

In the following table, *ARI* is used to be the evaluation metric. In Tab. 5, *CUBOS* is ranking first on all datasets whose effectiveness significantly surpasses that of other indices. *CCI* is the second best index. And *IE* is the worst index that is low-ranking on all datasets.

With respect to *F* (see Tab. 6), the effectiveness of *CUBOS* is still first-rate. Meanwhile *CCI* is slightly worse than *CUBOS* and *IE* is the worst index.

We now focus on the results shown in Tab. 7. *CUBOS* obtains the best evaluation results with *M* as metric. Nevertheless, *CUBOS* and *CCI* both perform best on these five datasets when *M* is used to evaluate the indices' effectiveness. Additionally, IE is still the worst performing index whose best ranking is only fifth place.

Next, we focus on the Tab. 8. *CUBOS* is also ranking first on all datasets with *NMI* as metric. And *CU* is ranking first on four datasets, *CCI* and *NCC* perform best on three datasets. The performance of *IE* is the worst.

With regard to *P*, Tab. 9 shows that *CUBOS*, *IE* and *CU* are all ranking first three times, and CCI performs poorly on all datasets. Specifically, the performance of *CUBOS* on the remaining two datasets which are not ranking first is better than that of *IE* and *CU* on their remaining two datasets.

Finally, Tab. 10 shows the evaluation results with *RI* as metric. The performance of *CUBOS* is excellent compared with other indices. *CCI* and *CU* are the second best indices. And *IE* is the worst index.

Table 4 Evaluation of all indices with A as metric

A	*CUBOS*	*CCI*	*CDCS*	*IE*	*CU*	*NCC*
BC	**(1)0.787**	(5)0.776	**(1)0.787**	(4)0.783	**(1)0.787**	(4)0.783
D	(2)0.831	(4)0.768	(6)0.639	**(1)0.915**	(4)0.768	(3)0.795
MB	**(1)0.877**	(6)0.774	**(1)0.877**	(4)0.858	**(1)0.877**	(4)0.858
S	**(1)1**	(5)0.787	(5)0.787	**(1)1**	**(1)1**	**(1)1**
CVR	(3)0.879	(3)0.879	(2)0.940	**(1)0.944**	(3)0.879	(3)0.879

Table 5 Evaluation of all indices with ARI as metric

ARI	*CUBOS*	*CCI*	*CDCS*	*IE*	*CU*	*NCC*
BC	**(1)0.247**	**(1)0.247**	**(1)0.247**	(6)0.041	(4)0.216	(5)0.105
D	**(1)0.694**	(3)0.678	(6)0.438	(5)0.496	(2)0.678	(3)0.678
MB	**(1)0.293**	**(1)0.293**	(3)0.218	(5)0.119	(4)0.213	(5)0.1186
S	**(1)1**	(6)0.654	(4)0.710	(4)0.960	**(1)1**	**(1)1**
CVR	**(1)0.574**	**(1)0.574**	(5)0.525	(6)0.204	**(1)0.574**	**(1)0.574**

Table 6 Evaluation of all indices with F as metric

F	*CUBOS*	*CCI*	*CDCS*	*IE*	*CU*	*NCC*
BC	**(1)0.755**	**(1)0.755**	**(1)0.755**	(6)0.346	(4)0.741	(5)0.564
D	**(1)0.831**	(3)0.797	(6)0.667	(5)0.688	(3)0.797	(2)0.800
MB	**(1)0.772**	**(1)0.772**	(3)0.632	(5)0.417	(4)0.627	(5)0.417
S	**(1)1**	(5)0.820	(5)0.820	(4)0.979	**(1)1**	**(1)1**
CVR	**(1)0.879**	**(1)0.879**	(5)0.830	(6)0.551	**(1)0.879**	**(1)0.879**

Table 7 Evaluation of all indices with M as metric

M	*CUBOS*	*CCI*	*CDCS*	*IE*	*CU*	*NCC*
BC	**(1)0.848**	**(1)0.848**	**(1)0.848**	(6)0.242	(4)0.823	(5)0.495
D	**(1)1**	**(1)1**	**(1)1**	(6)0.571	**(1)1**	**(1)1**
MB	**(1)0.774**	**(1)0.774**	(3)0.528	(5)0.274	(4)0.519	(5)0.274
S	**(1)1**	**(1)1**	**(1)1**	(6)0.979	**(1)1**	**(1)1**
CVR	**(1)0.879**	**(1)0.879**	(5)0.797	(6)0.392	**(1)0.879**	**(1)0.879**

Table 8 Evaluation of all indices with NMI as metric

NMI	*CUBOS*	*CCI*	*CDCS*	*IE*	*CU*	*NCC*
BC	**(1)0.144**	**(1)0.144**	**(1)0.144**	(5)0.101	(4)0.133	(6)0.097
D	**(1)0.777**	**(1)0.777**	(6)0.608	(5)0.672	**(1)0.777**	**(1)0.777**
MB	**(1)0.326**	(6)0.243	**(1)0.326**	(4)0.281	**(1)0.326**	(4)0.281
S	**(1)1**	(5)0.849	(5)0.849	(4)0.963	**(1)1**	**(1)1**
CVR	**(1)0.510**	**(1)0.510**	(5)0.505	(6)0.411	**(1)0.510**	**(1)0.510**

Table 9 Evaluation of all indices with P as metric

P	*CUBOS*	*CCI*	*CDCS*	*IE*	*CU*	*NCC*
BC	**(1)0.787**	(6)0.776	**(1)0.787**	(4)0.783	**(1)0.787**	(4)0.783
D	(2)0.831	(4)0.768	(6)0.639	**(1)0.915**	(4)0.768	(3)0.795
MB	**(1)0.877**	(6)0.774	**(1)0.877**	(4)0.858	**(1)0.877**	(4)0.858
S	**(1)1**	(5)0.787	(5)0.787	**(1)1**	**(1)1**	**(1)1**
CVR	(3)0.879	(3)0.879	(2)0.940	**(1)0.944**	(3)0.879	(3)0.879

Table 10 Evaluation of all indices with RI as metric

RI	*CUBOS*	*CCI*	*CDCS*	*IE*	*CU*	*NCC*
BC	**(1)0.651**	**(1)0.651**	**(1)0.651**	(6)0.454	(4)0.636	(5)0.519
D	**(1)0.890**	(4)0.879	(6)0.766	(5)0.872	**(1)0.890**	(3)0.882
MB	**(1)0.646**	**(1)0.646**	(3)0.611	(5)0.563	(4)0.609	(5)0.563
S	**(1)1**	(6)0.843	(5)0.880	(4)0.985	**(1)1**	**(1)1**
CVR	**(1)0.787**	**(1)0.787**	(5)0.763	(6)0.602	**(1)0.787**	**(1)0.787**

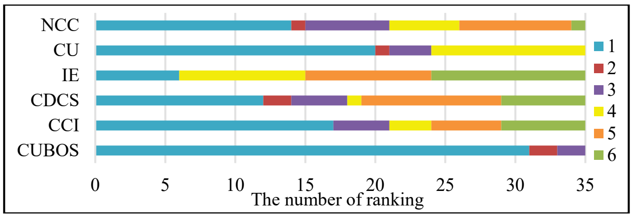

Figure 2 The number of ranking of internal cluster validity indices compared

From the above comprehensive analysis, it is clear that CUBOS can always choose a better clustering partition, compared with other internal indices for categorical data, no matter which external index is used.

Furthermore, we count the occurrence number of each ranking for each internal index compared in the experiments as shown in Fig. 2. It can be seen that *CUBOS* ranks first most frequently and its worst ranking is third, besides, *CUBOS* is ranking second and third on just a few datasets. Therefore, we could know that the performance of *CUBOS* proposed in this paper is significantly superior than other indices.

5 CONCLUSION

In this paper, we present a new internal cluster validity index for categorical data named *CUBOS*, in which a distance metric for categorical data *IDC* is derived from the Category Distance in an existing work and the paradigm of *S* index is used to construct *CUBOS*. The proposed index considers the relationship between different categorical attribute values and measures the distance between two categorical data objects exactly. Furthermore, the paradigms of *S* index and *IDC* are combined, so that much more detailed distribution information in clustering results of categorical data is explored and more precise evaluation results can be obtained. Experimental results on several *UCI* datasets show that *CUBOS* outperforms other internal cluster validity indices for categorical data compared. That demonstrates a reliable performance of our index and promises wide applicability in practice.

Acknowledgements

This work is supported by National Natural Science Foundation of China (NSFC) under Grant No.71271027.

6 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] Wen, L., Zhou, K., & Yang, S. (2019). A shape-based clustering method for pattern recognition of residential electricity consumption. *Journal of Cleaner Production, 212*(2019), 475-488. https://doi.org/10.1016/j.jclepro.2018.12.067

[3] Khalili-Damghani, K., Abdi, F., & Abolmakarem, S. (2018). Hybrid soft computing approach based on clustering, rule mining, and decision tree analysis for customer segmentation problem: Real case of customer-centric industries. *Applied Soft Computing, 73*(2018), 816-828. https://doi.org/10.1016/j.asoc.2018.09.001

[4] Faroughi, A. & Javidan, R. (2018). CANF: Clustering and anomaly detection method using nearest and farthest neighbour. *Future Generation Computer Systems, 89*(2018), 166-177. https://doi.org/10.1016/j.future.2018.06.031

[5] Zhang, D., Jin, D., Gong, Y., Chen, S., & Wang, C. (2015). Research of alarm correlations based on static defect detection. *Tehnicki Vjesnik-Technical Gazette, 22*(2), 311-318. https://doi.org/10.17559/TV-20150317102804

[6] Hachaj, T. & Ogiela, M. (2017). Clustering of trending topics in microblogging posts: A graph-based approach. *Future Generation Computer Systems, 67*(2017), 297-304. https://doi.org/10.1016/j.future.2016.04.009

[7] Lee, S., Jeong, Y., Kim, J., & Jeong, M. (2018). A new clustering validity index for arbitrary shape of clusters. *Pattern Recognition Letters, 112*(2018), 263-269. https://doi.org/10.1016/j.patrec.2018.08.005

[8] Dunn, J. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybernetics, 3*(3), 32-57. https://doi.org/10.1080/01969727308546046

[9] Calinski, T. & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, 3(1), 1-27. https://doi.org/10.1080/03610927408827101

[10] Liu, Y., Gao, X., Guo, H., & Wu, S. (2011). Ensembling clustering validation indices. *Computer Engineering and Applications*, 47(19), 15-17. https://doi:10.3778/j.issn.1002-8331.2011.19.005

[11] Davies, D. & Bouldin, D. (1979). A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence, 2*(1979), 224-227. https://doi.org/10.1109/TPAMI.1979.4766909

[12] Rousseeuw, P. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*(1987), 53-65. https://doi.org/10.1016/0377-0427(87)90125-7

[13] 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

[14] Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data mining and knowledge discovery, 2*(3), 283-304.

[15] Barbara, D., Li, Y., & Couto, J. (2002). COOLCAT: an entropy-based algorithm for categorical clustering. In *11th International conference on Information and knowledge management (CIKM)*, 582-589. https://doi.org/10.1145/584792.584888

[16] Gao, C., Pedrycz, W., & Miao, D. (2013). Rough subspace-based clustering ensemble for categorical data. *Soft Computing*, 17(9), 1643-1658. https://doi.org/10.1007/s00500-012-0972-8

[17] Chang, C. & Ding, Z. (2005). Categorical data visualization and clustering using subjective factors. *Data and Knowledge Engineering, 53*(3), 243-262. https://doi.org/10.1007/978-3-540-30076-2_23

[18] Gluck, M. (1985). Information, uncertainty, and the utility of categories. In *1985 Annual Conference of the Cognitive Science Society*, 283-287.

[19] Pierre, M. (1997). Clustering techniques. *Future Generation Computer Systems, 13*(1997), 135-147. https://doi.org/10.1016/s0167-739x(97)00017-4

[20] Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J., & Perona, I. (2010). An extensive comparative study of cluster validity indices. *Pattern Recognition, 46*(1), 243-256. https://doi.org/10.1016/j.patcog.2012.07.021

[21] Chen, B. & Yin, H. (2018). Learning category distance metric for data clustering. *Neurocomputing*, 306(2018), 160-170. https://doi.org/10.1016/j.neucom.2018.03.048

[22] Goodall, D. (1966). A new similarity index based on probability. *Biometrics, 22*(4), 882-907. https://doi.org/10.2307/2528080

[23] Wu, S., Feng, X., &Zhou, W. (2014). Spectral clustering of high-dimensional data exploiting sparse representation vectors. *Neurocomputing, 135*(2014), 229-239. https://doi.org/10.1016/j.neucom.2013.12.027

[24] Lei, Y., Bezdek, J., Romano, S., Vinh, N., 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

Contact information:

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

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