EAI Endorsed Transactions
on Collaborative Computing

Research Article EAI.EU

Analysis and improvement of evaluation indexes for clustering results

Hao Zhong[1,*], Huibing Zhang[2], Fei Jia[2]

[1]School of Computer Science, South China Normal University, Guangzhou 510631, China
[2]Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin 541004, China

Abstract

Clustering algorithm is the main field in collaborative computing of social network. How to evaluate clustering results accurately has become a hot spot in clustering algorithm research. Commonly used evaluation indexes are SC, DBI and CHI. There are two shortcomings in the calculation of three indexes. (1) Keep the number of clusters and the objects in the cluster unchanged. When transforming the feature vector, the three indexes will change greatly; (2) Keep the feature vector and the number of clusters unchanged. When changing the objects in the cluster, the three indexes will change tinily. This shows that the three indexes unable to evaluate the clustering results very well. Therefore, based on the calculation process of the three indexes, the paper proposes new three indexes - NSC, NDBI and NCHI. Through testing on standard data sets, three new indexes can better evaluate clustering results.

Received on 07 October 2019, 18 February 2020, published on 18 February 2020

Keywords: evaluation indexes, Calinski-Harabasz Index, Davies-Bouldin Index, Silhouette Coefficient

doi:10.4108/eai.9-10-2017.163211

1. Introduction

Clustering is an important algorithm for data mining in collaborative computing of social network. The purpose of clustering is to bring together similar objects and separate dissimilar objects. Since the object facing the clustering algorithm is unlabeled, the cluster to which the object belongs is unknown, so how to better evaluate the clustering result has become one of the research hotspots in unsupervised learning field. At present, there are three main indexes for the evaluation of unlabeled clustering results: Calinski-Harabasz Index[1], Davies-Bouldin Index[2] and Silhouette Coefficient[3].They define the calculation methods of intra-cluster relations and inter-cluster relations respectively, and evaluate the clustering results according to the combination of intra-cluster relations and inter-cluster relations. Because the three indexes can evaluate the clustering results intuitively, they can be applied to a wide range of clustering scenarios, or test the optimization effect of clustering algorithm.

Clustering tasks are divided into three steps. Firstly, objects are mapped to feature vectors with certain rules. Secondly, feature vectors are clustered by various clustering algorithms. Finally, CHI(Calinski-Harabasz Index), DBI(Davies-Bouldin Index) and SC(Silhouette Coefficient) are used to evaluate the clustering results. But there are two problems in evaluating the clustering results: (1) Different researchers will propose different rules for generating feature vectors. For the same set of objects, different feature vectors will be generated by using different generation rules. Keep the number of clusters and the objects in each cluster unchanged. When the vector elements change, the calculated values of CHI, DBI and SC will change greatly. (2) Different researchers will propose different clustering algorithms. Keep the number of clusters and the elements of each vector unchanged. When the objects in each cluster change, the calculated values of CHI, DBI and SC will change tinily. In order to solve the above two problems, and make the clustering results better evaluated, based on the calculation process of three indexes, this paper proposes new indexes NSC(New SC), NDBI(New DBI) and NCHI(New CHI). The main contributions are as follows:

- Keep the number of clusters and the objects in each cluster unchanged. When the feature vectors change, the problem that the calculated values of indexes change greatly is solved to some extent. In calculating the relationship between feature vectors, the cosine of the angle between vectors

*Corresponding author. Email: scnuzhonghao@foxmail.com

is used instead of the distance between vectors. When the feature vectors change, the calculated value of the relationship between the feature vectors will only change in the $[0,1]$ interval. Thus, the changes of the three indexes can be stabilized in a small interval.

- Keep the number of clusters and the elements of each vector unchanged. When the objects in the clusters change, the problem that the calculated values of indexes change tinily is solved to some extent. In calculating the relationship between feature vectors, taking the number of elements in each cluster as a coefficient, expand the calculation values of inter-cluster relations and intra-cluster relations, but the range of increase of the two calculated values is different. Thus, the changes of the three indexes are more obvious.

2. Related Works

The indexes CHI, DBI and SC can be applied to evaluate the cluster effect in various scenarios. For example, Hassani et al.[4] believed that when clustering algorithm is applied to analyze dynamic unlabeled data, it is impossible to use labels to evaluate clustering results. CHI is an evaluation index based on data itself, is suitable for evaluating clustering results of dynamic data. Schkafer et al.[5] analyzed the dynamic consumption data and static information data of users, clustered users by using hybrid fuzzy clustering algorithm, CHI, DBI and SC are used to evaluate the results of user clustering. In order to realize the division and management of different regions, Arroyo et al.[6] analyzed the meteorological data of Spain, used K-Means and other clustering algorithms to cluster the regions, and used CHI, DBI and SC to evaluate the clustering results. Babichev et al.[7] took the gene expression sequence of cancer patients as a feature, used different criteria to measure the similarity between the features, clustered the cancer patients according to the similarity, and used CHI index to evaluate the clustering results. In order to manage power resources better, Damayanti et al.[8] analyzed power consumption in each time period, applied clustering algorithm in every time period, and selected the best clustering result according to DBI index. In order to improve the accuracy of software size estimation, Prokopov et al.[9] proposed a new clustering method based on use case points, Compared with clustering algorithms such as K-Means, the evaluation of clustering results using indexes such as CHI and SC. Umam et al.[10] proposed a hybrid clustering method based on K-Means clustering and hierarchical clustering. The sequence of DNA was used as a feature to cluster, the DBI was used to evaluate clustering results. In order to achieve better management of video data, Kumar et al.[11] proposed an equal-partition clustering technology, achieved video clustering in real-time applications, using the DBI index to evaluate the results of video clustering. In order to achieve clustering of hospital patients, based on algebraic structure, Thanh et al.[12] constructed a neutron recommendation equivalent matrix for hospital patients, and performed λ-cutting on the matrix. The result of the cutting is the clustering result of the patient, using DBI to evaluate patient clustering results. In order to place public facilities in the appropriate population, Kisore et al.[13] proposed a clustering algorithm base on generalized density. Users were clustered based on their requirements, preferences and geographical location. SC and DBI were used to evaluate the results of user clustering. In order to realize the analysis of telecommunication service cluster, Tosida et al.[14] proposed a self-organizing mapping algorithm based on artificial neural network system, clustered telecommunication service facilities, and used DBI as the evaluation index of clustering results. In order to provide data for the construction of wind power plants, according to the wind speed in Turkey, Yesilbudak et al.[15] used K-Means clustering algorithm to divide the regions, and selected the best number of clusters according to the change of SC. In order to recommend different types of movies to everyone, Alfarizy et al.[16] extracts the valid information in the subtitles of movies, and clusters the movies based on these valid information. The default clustering effect is the best when SC calculates the maximum value.Rani et al.[17] applied clustering algorithm to news text clustering, extract the words in news text as features. Hierarchical clustering and K-Means clustering are used to get different clustering results, and the optimal clustering results are selected according to the change of SC. Sarasa et al.[18] clustered birds and animals based on audio data, when measuring the similarity between audio features, a method of normalized compressed distance was proposed. Hierarchical clustering was used to cluster audio features, SC was used to evaluate the clustering results. Mago et al.[19] applied clustering algorithm to neonatal clustering. According to the characteristic data of neonates, hierarchical clustering was used to cluster neonates, and SC was used to evaluate the clustering results, which could provide data for doctors to diagnose neonates with different conditions.

The indexes CHI, DBI and SC can also evaluate the clustering algorithm. For example, Raposo et al.[20] proposed an automatic clustering algorithm based on genetic algorithm, which can select the best number of clustering in the data set, and used CHI as the evaluation index of clustering results, compared with K-Means and fuzzy C-means algorithm. In order to realize the clustering of hospital patients, Li et al.[21] proposed

a multi-objective clustering algorithm. Based on multi-objective differential evolution algorithm, optimized multiple objective at the same time, finded the optimal clustering result. Siddiqi et al.[22] proposed using a heuristic algorithm to solve the clustering problem. The algorithm was divided into two parts. The first part was using the greedy algorithm to select the data points with higher resolution as the cluster center. The second part set CHI as the objective function, when the objective function takes the maximum value, obtains the optimal clustering result. In order to achieve image clustering, Toz et al.[23] combined fuzzy C-means clustering algorithm with backtracking search optimization algorithm, improved the local search ability of the algorithm by optimizing the objective function, and used the DBI to evaluate the clustering results. Gowri et al.[24] believe that when clustering large data, it is necessary to use MapReduce framework for distributed processing of data. Under the premise of not changing the clustering result, according to the change of DBI, it can be proved that using the MapReduce framework to preprocess the data can shorten the running time of the clustering algorithm. Andryani et al.[25] applied the fuzzy C-means clustering algorithm to the clustering task of DNA, and proposed combining the splitting algorithm with the fuzzy C-means clustering algorithm. The process of solving the minimum value of the objective function can be simplified, the final clustering results were evaluated by using DBI. Halim et al.[26] transformed the data into a probability map representation, and proposed a clustering method based on the density, and then completed the clustering of the data. The clustering results were evaluated by using the DBI and SC. Ketsuwan et al.[27] proposed that using linear discriminant analysis to reduce the dimension of feature vectors, it can further minimize intra-cluster dispersion and maximize inter-cluster separation, thus improving the DBI of clustering results and obtaining better clustering results. Hasanzadeh et al.[28] proposed an automatic learning machine clustering algorithm. In the clustering process, the reinforcement signal was defined according to the Euclidean distance between the data points and the clustering centers. The automatic learning machine judged the cluster of each data point by correcting the reinforcement signal. SC was used to compare the clustering results with K-Means.

Aiming at the problems existing in the calculation of CHI, DBI and SC, Amorim et al.[29] believed that scaling features would affect the indexes such as CHI and SC. In order to improve the indexes of clustering, data sets were indexed according to such indexes as CHI and SC, it can find the ideal scaling factors. Fernandez et al.[30] believed that constructing different features on the same object would affect the CHI. Therefore, based on Laplace value and CHI, a feature selection framework was proposed, which is more suitable for object construction. The characteristics of clustering tasks achieve better clustering results. Cheng et al.[31] considered that SC and DBI could not be applied to evaluate the clustering results of all types of data. Therefore, based on SC, a new metric was proposed — the clustering effectiveness based on local clustering centers. The index selects the local cluster center with local maximum density as the representative point, which can more accurately evaluate the difference between the cluster centers.

In the above research, when applying CHI, DBI and SC to evaluate clustering results, some researchers have suggested that there are shortcomings in the calculation process. But only used CHI, DBI and SC as objective functions, by searching for the optimal solution of the objective function. constructed the feature or to select the number of clusters, and the calculation process of CHI, DBI and SC was not changed. The paper will analyze the problems in calculations process of CHI, DBI and SC, and propose NCHI, NDBI and NSC as new indexes. so that these new indexes can better evaluate the clustering results.

3. Analysis of Evaluation Indexes

3.1. Definition of Relevant Symbols

According to the definition of three indexes, the variables and functions used in the calculation of indexes are defined in Table 1.

Table 1. Symbols used in the paper

Feature vectors	$X = \{X_1, X_2, X_3, ..., X_n\}$
Cluster centers	$C = \{C_1, C_2, C_3, ..., C_k\}$
Feature vectors in cluster k	$K = \{K_1, K_2, K_3, ..., K_m\}$
Elements of vector X_n	$X_n = \{X_{n,1}, X_{n,2}, X_{n,3}, ..., X_{n,d}\}$
Euclidean distance of X_i and X_j	$\mathrm{Eudis}(X_i, X_j)$
Number of elements in set S	$\mathrm{Len}(S)$
Minimum element in set S	$\mathrm{Min}(S)$
Maximum element in set S	$\mathrm{Max}(S)$
Average value of all elements in set S	$\mathrm{Mean}(S)$
Cosine distance of X_i and X_j	$\mathrm{Cos}(X_i, X_j)$

3.2. Calinski-Harabasz Index

When evaluating the clustering results, the larger the calculated value of the CHI (Calinski-Harabasz Index), the better the clustering effect. CHI contains the following specific meanings. There are two concepts: group dispersion and within-cluster dispersion. For cluster k, the group dispersion represented by $\mathrm{Grd}(k)$, the within-cluster dispersion represented by $\mathrm{Wcd}(k)$. The calculation is shown in formulas (1) and (2).

$$Grd(k) = Len(K) * Eudis(Mean(X), C_k)^2 \quad (1)$$

$$Wcd(k) = \sum_{j=1}^{m} Eudis(C_k, K_j)^2 \quad (2)$$

According to formulas (1) and (2), the calculation of CHI is shown in formula (3).

$$CHI = \frac{\sum_{i=1}^{k} Grd(i)}{\sum_{i=1}^{k} Wcd(i)} * \frac{n-k}{k-1} \quad (3)$$

3.3. Davies-Bouldin Index

When evaluating the clustering results, the smaller the calculated value of the DBI (Davies-Bouldin Index), the better the clustering effect. DBI contains the following specific meanings. There is one concept: average similarity between two clusters. The similarity between cluster i and cluster j represented by Sim(i,j). Formula (2) is included in the calculation process, as shown in formula (4).

$$Sim(i,j) = \frac{\frac{Wcd(i)}{Len(I)} + \frac{Wcd(j)}{Len(J)}}{Eudis(C_i, C_j)^2} \quad (4)$$

According to formula (4), the similarity between cluster i and any other cluster is calculated, the similarity list $SL(i)$ of cluster i is constructed. $SL(i)$=[Sim(i,1), Sim(i,2), Sim(i,3), ..., Sim(i,k)], extract the maximum value in the list. The maximum values in the similarity list of all clusters are added together, and the mean values are calculated as DBI, as shown in formula (5):

$$DBI = \frac{1}{k} * \sum_{i=1}^{k} Max(SL(i)) \quad (5)$$

3.4. Silhouette Coefficient

When evaluating the clustering results, the larger the calculated value of the SC (Silhouette Coefficient), the better the clustering effect. SC contains the following specific meanings. There are two concepts: the mean distance between a sample and all other points in the same cluster, and the mean distance between a sample and all other points in the next nearest cluster. For a sample K_m, the mean distance between it and all other points in the same cluster represented by Wcmd(K_m), the mean distance between it and all points in other cluster k' represented by Gmd(K_m,k'). The calculation is shown in formulas (6) and (7).

$$Wcmd(K_m) = \frac{1}{m} * \sum_{i=1}^{m} Eudis(K_m, K_i) \quad (6)$$

$$Gmd(K_m, k') = \frac{1}{Len(K')} * \sum_{i=1}^{m'} Eudis(K_m, K_i') \quad (7)$$

According to formula (7), get a list of the average distances between K'_m and other clusters GD(K'_m). GD(K'_m) = [Gmd(K'$_m$, 1), Gmd(K'$_m$, 2), ..., Gmd(K'$_m$, k'-1), Gmd(K'$_m$, k'+1), Gmd(K'$_m$, k)]. The minimum value of GD(K'_m) is extracted, combined with formula (6), SC is calculated as shown in equation (8).

$$SC = \frac{1}{n} * \sum_{k'=1}^{k} \sum_{i=1}^{m} \frac{Min(GD(K_i')) - Wcmd(K_i')}{Max(Min(GD(K_i')), Wcmd(K_i'))} \quad (8)$$

3.5. Problem Description

Keep the number of clusters and the objects in each cluster unchanged, then Len(K), k and n remain unchanged. According to the formula of Eudis(), when the vector element K_m changes in interval $[0, +\infty]$, the Grd(K_m), Wck(K_m), Sim(K_m), Gmd(K_m, k') and Wcmd(K_m) related to Eudis() all change in interval $[0, +\infty]$. Finally, the calculated values of CHI, DBI and SC will change, as shown in formula (9).

$$\begin{cases} Eudis(Mean(X), C_k) \in [0,+\infty], & K_m \in [0,+\infty]. \\ Eudis(C_k, K_j) \in [0,+\infty], & K_m \in [0,+\infty]. \\ Eudis(C_i, C_j) \in [0,+\infty], & K_m \in [0,+\infty]. \\ Eudis(K_m, K_m') \in [0,+\infty], & K_m \in [0,+\infty]. \end{cases} \Longrightarrow \begin{cases} Grd(k) \in [0,+\infty], & K_m \in [0,+\infty]. \\ Wcd(k) \in [0,+\infty], & K_m \in [0,+\infty]. \\ Sim(i,j) \in [0,+\infty], & K_m \in [0,+\infty]. \\ Gmd(K_m, k') \in [0,+\infty], & K_m \in [0,+\infty]. \\ Wcmd(K_m) \in [0,+\infty], & K_m \in [0,+\infty]. \end{cases} \Longrightarrow \begin{cases} CHI \in [0,+\infty], & K_m \in [0,+\infty]. \\ DBI \in [0,+\infty], & K_m \in [0,+\infty]. \\ SC \in [0,+\infty], & K_m \in [0,+\infty]. \end{cases} \quad (9)$$

If logarithmic or exponential transformations are performed on vector elements, the variation rates of Grd(K_m), Wck(K_m), Sim(K_m), Gmd(K_m, k') and Wcmd(K_m) are different. CHI, DBI and SC will have maximum and minimum values in the interval $[0, +\infty]$ of K_m. It is impossible to evaluate the clustering results from the calculated values of the indexes.

Keep the number of clusters and the elements of the feature vector unchanged, when the objects in the cluster change, Mean(X), k and n remain unchanged. The result of clustering algorithm is that similar feature vectors are clustered into one cluster. Under the premise of the same number of clusters, Assuming that the set of objects in cluster k changes, a new clustering center C_k' is obtained. The new clustering

center C_k' is approximately equal to the original clustering center C_k, that is, $C_k \approx C_k$'. In particular, CHI, DBI and SC all use the average distance method to describe the relationships between and within clusters, which further narrows the gap between the relationships between and within clusters. Therefore, the phenomenon of object change in clustering results can not be well reflected.

4. Redefinition of Evaluation Indexes

4.1. Solutions to problems

When the element of feature vectors changes in interval $[0, +\infty]$, if Eudis() is used to represent the relationship between any two feature vectors, the calculated value of Eudis() will change in interval interval $[0, +\infty]$ equally. The paper proposes to use Cos() to measure the relationship between vectors. When the vector elements change in interval $[0, +\infty]$, the calculated value of Cos() will be stable in interval $[0, 1]$. Suppose that the relationship between vectors a and b is calculated by Cos(), as shown in formula (10).

$$Cos(a, b) = \left| \frac{a \cdot b}{|a||b|} \right| \tag{10}$$

When the object set in the cluster changes, the number of elements in the set changes relatively. Therefore, when calculating inter-cluster or intra-cluster relationships, the number of elements in the cluster is used as a coefficient, to increase the calculated value of intra-cluster or intra-cluster relationship. Thereby expanding the calculated values of CHI, DBI and SC. The calculated values of CHI, DBI and SC increase in magnitude, which can better reflect the change of object sets in clusters.

4.2. Redefinition of Calinski-Harabasz Index

According to the relevant definition of CHI, the calculation process Eudis() existing in Grd(k) and Wcd(k) is replaced by Cos(). Moreover, in the calculation process of Grd(k), Len(K) is used as a coefficient to expand the calculated value of Grd(k), thereby increasing the gap between Grd(k) and Wcd(k). The specific calculation processes of Grd'(k) and Wcd'(k) are shown in formulas (11) and (12).

$$Grd'(k) = Len(K)^2 * Cos(Mean(X), C_k) \tag{11}$$

$$Wcd'(k) = \sum_{j=1}^{m} Cos(C_k, K_j) \tag{12}$$

The larger the calculated value of CHI, the better the clustering result. Without changing this principle, the NCHI calculation defined is shown in formula (13).

$$NCHI = \frac{\sum_{i=1}^{k} Grd'(i)}{\sum_{i=1}^{k} Wcd'(i)} \tag{13}$$

4.3. Redefinition of Davies-Bouldin Index

According to the relevant definition of DBI, the calculation process Eudis() existing in Wcd(k) and Sim(i,j) is replaced by Cos(). Moreover, in the calculation process of Sim(i,j), Len(I) and Len(J) are used as a coefficient to expand the calculated value of Wcd'(i) and Wcd'(j), thereby increasing the gap between the numerator and the denominator in the formula of DBI. The specific calculation processes of Sim'(i,j) is shown in formulas (14).

$$Sim'(i, j) = \frac{Len(I) * Wcd'(i) + Len(J) * Wcd'(j)}{Cos(C_i, C_j)} \tag{14}$$

The smaller the calculated value of DBI, the better the clustering result. Without changing this principle, the similarity between cluster i and any other cluster is calculated, the similarity list $SL'(i)$ of cluster i is constructed. $SL'(i)$=[Sim'(i,1), Sim'(i,2), Sim'(i,3), ..., Sim'(i,k)], extract the maximum value in the list. The maximum values in the similarity list of all clusters are added together, and the mean values are calculated as NDBI, as shown in formula (15):

$$NDBI = \frac{1}{k} * \sum_{i=1}^{k} Max(SL'(i)) \tag{15}$$

4.4. Redefinition of Silhouette Coefficient

According to the relevant definition of SC, the calculation process Eudis() existing in Gmd(K_m,k') and Wcmd(K_m) is replaced by Cos(). Moreover, in the calculation process of Gmd(K_m,k') and Wcmd(K_m), Len(K') is used as a coefficient to expand the calculated value of Gmd(K_m,k'), Len(K) is used as a coefficient to expand the calculated value of Wcmd(K_m). Thereby increasing the gap between Gmd(K_m,k') and Wcmd(K_m) in the formula of SC. The specific calculation processes of Wcmd'(K_m) and Gmd'(K_m,k') are shown in formulas (16) and (17).

$$Wcmd'(K_m) = \sum_{i=1}^{m} Cos(K_m, K_i) \tag{16}$$

$$Gmd'(K_m, k') = \sum_{i=1}^{m'} Cos(K_m, K_i') \tag{17}$$

The larger the calculated value of SC, the better the clustering result. Without changing this principle, the average distances between K'_m and other clusters

is calculated, the distance list GD'(K'_m) of K'_m is constructed. GD'(K'_m) = [Gmd'(K'_m, 1), Gmd'(K'_m, 2), ..., Gmd'(K'_m, k'-1), Gmd'(K'_m, k'+1), Gmd'(K'_m, k)], extract the minimum value in the list. The NSC calculation defined is shown in formula (18).

$$SC = \frac{1}{n} * \sum_{k'=1}^{k} \sum_{i=1}^{m} \frac{Min(GD'(K'_i))}{Wcmd'(K'_i)} \tag{18}$$

5. Data sets and Experiments

5.1. Data Set and Evaluation Standard

In the sklearn database, four standard data sets are selected to test the effectiveness of the proposed method. The data set is suitable for classification tasks. The selected data set and the invocation method are shown in Table 2.

Table 2. Datasets

Invoke Method	Representation
datasets.load-iris()	iris
datasets.load-breast-cancer()	breast
datasets.load-digits()	digits
datasets.load-wine()	wine

After redefining the three indexes, the dimensions of the three indexes changed. Therefore, the paper used the CV(coefficient of variation) as the evaluation standard of the experimental results. The calculated value of CV indicates the degree of dispersion of the original index and the new index. The larger the absolute value of the CV, the more discrete the calculated value of the index. The calculation method is as shown in formula (19):

$$CV = \frac{\frac{1}{N-1} * (\sum_{i=1}^{N} (EI_i - \frac{1}{N} * \sum_{j=1}^{N} EI_j)^2)}{\frac{1}{N} * \sum_{i=1}^{N} EI_i} \tag{19}$$

Among them, N represents the number of samples (or experiments) and EI represents the calculation results of indexes.

5.2. An Example of Calculating Indexes

In the first experiment of this paper, an example is given to prove the problems of indexes in calculation. Data iris is used as test data, data iris is labeled data, and its label is defaulted to be the optimal clustering result. Keeping the clustering results unchanged, a negative exponential transformation (represented by New iris) is adopted for all vector elements in data iris, is $X_n = e^{-\frac{1}{X_n}}$. Taking the optimal clustering results, the CHI, DBI and SC are used to evaluate the original data, and to evaluate the transformed data. The experimental results are shown in Table 3:

Table 3. Clustering results of iris

dataset	CHI	DBI	SC
iris	487.33	0.75	0.50
New iris	1631.65	0.54	0.61

As shown in Table 3, only the vector elements are transformed, and the calculated values from CHI, DBI, and SC reflect the improvement of the clustering effect, but in fact the clustering results are not changed. It is proved that the calculated values of CHI, DBI and SC are sensitive to vector elements.

Similarly, data iris is used as test data. Because K-Means has the characteristics of randomly selecting the initial cluster center. The K-Means clustering algorithm is used to analyze the data iris twice. In the clustering process, the number of clusters of the two clustering results is the same, and the objects in each cluster are different. Two clustering results were evaluated using CHI, DBI, and SC, respectively. The experimental results are shown in Table 4:

Table 4. Clustering results of iris

dataset	CHI	DBI	SC
iris	437.60	0.92	0.36
New iris	438.43	0.95	0.35

As shown in Table 4, when evaluating the clustering results, although the objects in each cluster have changed, but the change of the calculated value of three indexes is very small, so the calculated values of the indexes do not significantly reflect the changes in the clustering results.

5.3. Clustering Results Testing-Changing Feature Vectors

In the second experiment, this paper assumes that the elements of each feature vector are transformed by negative exponential transformation, is $X_n = e^{-\frac{\alpha}{X_n}}$. α is the transformation coefficient. Set the change interval of α is $[1, 1000]$, and increase the number of α once, complete a cluster analysis. Each clustering result was evaluated using CHI, DBI, SC, NCHI, NDBI, and NSC. Calculate the CV of 1000 clustering results. The experimental results are shown in Table 5:

Table 5. Clustering results - Changing Feature Vectors

dataset	CHI	DBI	SC	NCHI	NDBI	NSC
iris	0.221	0.294	0.239	0.040	0.013	0.016
breast	0.350	-2.106	0.202	0.001	0.001	0.002
digits	0.049	0.055	0.030	0.010	0.004	0.042
wine	0.200	-0.245	0.101	0.002	0.001	0.003

When evaluating the same clustering results, the ideal situation is that the calculated values of the indexes are unchanged. As shown in Table 5, the CV of CHI, DBI, and SC are larger, and the CV of NCHI, NDBI, and NSC are smaller. This indicates that the calculated values of CHI, DBI, and SC vary more, and the calculated values of NCHI, NDBI, and NSC vary less. The experimental results show that, the calculated value of the new index is more stable when the feature is transformed, and the influence of the feature transform on the calculated value is reduced to some extent, and the clustering result can be better evaluated.

5.4. Clustering Results Testing–Changing Objects in Clusters

Since the K-Means clustering algorithm is random when selecting the initial cluster center, when the K-Means clustering algorithm is used to cluster the same data, the clustering results may be different each time. In the third experiment of the paper, the number of clusters is set to 8, and each data set is clustered 500 times. Each clustering result is evaluated by CHI, DBI, SC, NCHI, NDBI and NSC. The CV of 500 clustering results is calculated. The experimental results are shown in Table 6.

Table 6. Clustering results – Changing Objects in Clusters

dataset	CHI	DBI	SC	NCHI	NDBI	NSC
iris	0.010	0.024	0.030	0.065	0.022	0.345
breast	0.004	0.026	0.009	0.068	0.006	0.110
digits	0.007	0.025	0.010	0.063	0.011	0.099
wine	0.010	0.013	0.006	0.054	0.005	0.013

When the object in the cluster changes, under ideal situation, the calculated value of the indexes change significantly. As shown in Table 6, the CV of CHI, DBI, and SC are smaller, and the CV of NCHI, NDBI, and NSC are larger. This indicates that the calculated values of CHI, DBI, and SC vary less, and the calculated values of NCHI, NDBI, and NSC vary more. The experimental results show that, when the object in the cluster changes, the calculated value of the new index changes more obviously. It can more clearly show that the objects in the cluster have changed, and the clustering result can be better evaluated.

6. Conclusion

There are three main indexes for evaluating the clustering results. The paper analyzes two problems in the calculation of the three indexes and gives corresponding solutions. The first problem is that if the number of clusters is constant and the objects in the cluster are unchanged, the calculated value of indexes will change greatly when only the feature vector is changed. The cosine between the vectors is used instead of the distance between the vectors, which enables the three indexes to be stable within a small interval. The second problem is that the feature vectors are unchanged, the number of clusters is constant, and when the objects in the clusters are changed, the calculated value of the indexes does not change much. The paper uses the number of elements in each cluster as a coefficient, to expand the calculation between the inter-cluster relationship and the intra-cluster relationship, so that the three indexes can show a significant change. Two improvements are made to the indexes, so that the new indexes can better evaluate the clustering results, and the effectiveness of the improvement is proved by experiments.

In the future research work, it will be based on the following two aspects: Firstly, in an ideal situation, when evaluating the clustering results after feature transformation, the calculated values of the indexes should remain unchanged. After the improvement of the paper, there are still some fluctuations in the calculated values of the indexes, and future research work will try to further reduce the variation of the calculated values. Secondly, in an ideal situation, when evaluating the clustering results after the objects changed in the cluster, the calculated values of the indexes should change significantly. Through the improvement of the paper, the change range is only improved on the basis of the original indexes, and the future work will try to further increase the range of the change in the calculated value.

References

[1] Tadeusz Caliński and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27, 1974.

[2] David L Davies and Donald W Bouldin. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227, 1979.

[3] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[4] Marwan Hassani and Thomas Seidl. Using internal evaluation measures to validate the quality of diverse stream clustering algorithms. *Vietnam Journal of Computer Science*, 4(3):171–183, 2017.

[5] Hanna Schäfer, Joaquim L Viegas, Marta C Ferreira, Susana M Vieira, and João MC Sousa. Analysing the segmentation of energy consumers using mixed fuzzy clustering. In *2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–7. IEEE, 2015.

[6] Ángel Arroyo, Álvaro Herrero, Verónica Tricio, and Emilio Corchado. Analysis of meteorological conditions in spain by means of clustering techniques. *Journal of Applied Logic*, 24:76–89, 2017.

[7] S Babichev, Mohamed Ali Taif, V Lytvynenko, and V Osypenko. Criterial analysis of gene expression sequences to create the objective clustering inductive technology. In *2017 IEEE 37th International Conference on Electronics and Nanotechnology (ELNANO)*, pages 244–248. IEEE, 2017.

[8] R Damayanti, AG Abdullah, W Purnama, and ABD Nandiyanto. Electrical load profile analysis using clustering techniques. In *IOP Conference Series: Materials Science and Engineering*, volume 180, page 012081. IOP Publishing, 2017.

[9] Zdenka Prokopová, Radek Silhavy, and Petr Silhavy. The effects of clustering to software size estimation for the use case points methods. In *Computer Science On-line Conference*, pages 479–490. Springer, 2017.

[10] Khoirul Umam, Alhadi Bustamam, and Dian Lestari. Application of hybrid clustering using parallel k-means algorithm and diana algorithm. In *AIP Conference Proceedings*, volume 1825, page 020024. AIP Publishing, 2017.

[11] Krishan Kumar, Deepti D Shrimankar, and Navjot Singh. Equal partition based clustering approach for event summarization in videos. In *2016 12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, pages 119–126. IEEE, 2016.

[12] Nguyen Dang Thanh, Mumtaz Ali, et al. A novel clustering algorithm in a neutrosophic recommender system for medical diagnosis. *Cognitive Computation*, 9(4):526–544, 2017.

[13] N Raghu Kisore and CH B Koteswaraiah. Improving atm coverage area using density based clustering algorithm and voronoi diagrams. *Information Sciences*, 376:1–20, 2017.

[14] ET Tosida, S Maryana, H Thaheer, et al. Implementation of self organizing map (som) as decision support: Indonesian telematics services msmes empowerment. In *IOP Conference Series: Materials Science and Engineering*, volume 166, page 012017. IOP Publishing, 2017.

[15] Mehmet Yesilbudak. Clustering analysis of multidimensional wind speed data using k-means approach. In *2016 IEEE International Conference on Renewable Energy Research and Applications (ICRERA)*, pages 961–965. IEEE, 2016.

[16] AD Alfarizy, B Sartono, et al. Clustering box office movie with partition around medoids (pam) algorithm based on text mining of indonesian subtitle. In *IOP Conference Series: Earth and Environmental Science*, volume 58, page 012032. IOP Publishing, 2017.

[17] Usha Rani and Shashank Sahu. Comparison of clustering techniques for measuring similarity in articles. In *2017 3rd International Conference on Computational Intelligence & Communication Technology (CICT)*, pages 1–7. IEEE, 2017.

[18] Guillermo Sarasa, Ana Granados, and Francisco B Rodriguez. An approach of algorithmic clustering based on string compression to identify bird songs species in xeno-canto database. In *2017 3rd International Conference on Frontiers of Signal Processing (ICFSP)*, pages 101–104. IEEE, 2017.

[19] Nikhit Mago, Rudresh D Shirwaikar, U Dinesh Acharya, K Govardhan Hegde, Leslie Edward S Lewis, and M Shivakumar. Partition and hierarchical based clustering techniques for analysis of neonatal data. In *Proceedings of International Conference on Cognition and Recognition*, pages 345–355. Springer, 2018.

[20] Carolina Raposo, Carlos Henggeler Antunes, and Joao Pedro Barreto. Automatic clustering using a genetic algorithm with new solution encoding and operators. In *International Conference on Computational Science and Its Applications*, pages 92–103. Springer, 2014.

[21] Xiangtao Li and Ka-Chun Wong. Evolutionary multiobjective clustering and its applications to patient stratification. *IEEE transactions on cybernetics*, 49(5):1680–1693, 2018.

[22] Umair F Siddiqi and Sadiq M Sait. A new heuristic for the data clustering problem. *IEEE Access*, 5:6801–6812, 2017.

[23] Güliz Toz, İbrahim Yücedağ, and Pakize Erdoğmuş. A fuzzy image clustering method based on an improved backtracking search optimization algorithm with an inertia weight parameter. *Journal of King Saud University-Computer and Information Sciences*, 31(3):295–303, 2019.

[24] R Gowri and R Rathipriya. Quality based clustering using mapreduce framework. In *2016 Online International Conference on Green Engineering and Technologies (IC-GET)*, pages 1–5. IEEE, 2016.

[25] Diyah Septi Andryani, Alhadi Bustamam, and Dian Lestari. The implementation of hybrid clustering using fuzzy c–means and divisive algorithm for analyzing dna human papillomavirus cause of cervical cancer. In *AIP Conference Proceedings*, volume 1825, page 020003. AIP Publishing, 2017.

[26] Zahid Halim and Jamal Hussain Khattak. Density-based clustering of big probabilistic graphs. *Evolving Systems*, pages 1–18, 2018.

[27] Raywut Ketsuwan and Praisan Padungweang. A linear discriminant analysis using weighted local structure information. In *2017 14th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, pages 1–5. IEEE, 2017.

[28] Mohammad Hasanzadeh-Mofrad and Alireza Rezvanian. Learning automata clustering. *Journal of computational science*, 24:379–388, 2018.

[29] Renato Cordeiro de Amorim and Christian Hennig. Recovering the number of clusters in data sets with noise features using feature rescaling factors. *Information Sciences*, 324:126–145, 2015.

[30] Saúl Solorio-Fernández, J Ariel Carrasco-Ochoa, and José Fco Martínez-Trinidad. A new hybrid filter–wrapper feature selection method for clustering based on ranking. *Neurocomputing*, 214:866–880, 2016.

[31] Dongdong Cheng, Qingsheng Zhu, Jinlong Huang, Quanwang Wu, and Lijun Yang. A novel cluster validity index based on local cores. *IEEE transactions on neural networks and learning systems*, 30(4):985–999, 2018.