An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index

Xu Wang, Yusheng Xu
School of Information Science and Engineering, Lanzhou University, Lanzhou, Gansu,730000, China
wx2700317@163.com, xuyusheng@lzu.edu.cn

Abstract. The evaluation of clustering effects has been an important issue for a long time. How to effectively evaluate the clustering results of clustering algorithms is the key to the problem. The clustering effect evaluation is generally divided into internal clustering effect evaluation and external clustering effect evaluation. This paper focuses on the internal clustering effect evaluation, and proposes an improved index based on the Silhouette index and the Calinski-Harabasz index: Peak Weight Index (PWI). PWI combines the characteristics of Silhouette index and Calinski-Harabasz index, and takes the peak value of the two indexes as the impact point and gives appropriate weight within a certain range. Silhouette index and Calinski-Harabasz index will help improve the fluctuation of clustering results in the data set. Through the simulation experiments on four self-built influence data sets and two real data sets, it will prove that the PWI has excellent evaluation of clustering results.

1. Introduction
Clustering divides a group of unlabeled data into a series of subsets. The objects in the same subset are closely related, and the objects in different subsets are alienated. Clustering analyzes data through various constraints on a similar basis, and has a wide range of applications in mathematics, statistics, computer science, and economics [1][2]. Clustering is a very important unsupervised learning problem, and the clustering effect directly affects the analysis of clustering results. Finding an effective and accurate clustering effect evaluation algorithm is one of the difficult problems researchers have been trying to solve. The classical clustering algorithms include K-means, Hierarchical clustering, etc.

In 1974, Bezdek first proposed the validity index Partition Coefficient (PC) defined by the attribute[3]. The evaluation of clustering effects can be generally divided into the evaluation of external clustering effects and the evaluation of internal clustering effects. The evaluation of the external clustering effect is to refer to some reference objects, and the results are analyzed, usually based on the labels of the data itself. The Rand Index (RI) is the classic one of these evaluation indexes [4], since the Rand Index does not guarantee that the two clusters do not overlap when the category labels are randomly assigned, the Adjusted Rand Index (ARI) is proposed to improve the defect [5]. The index can be used to compare the similarity between clustering results of arbitrary clustering algorithms. There is also the Mutual Information (MI) [6], which measures the similarity of data distribution in the two data sets, and also proposes the improved index Adjusted Mutual Information (AMI). The Homogeneity and Completeness indexes attribute similar data to the same cluster. The same cluster contains only one type of sample, and the index V-measure is proposed based on the harmonic mean of the two. There are also other external indexes such as Fowlkes-Mallows Index (FMI) [7], Jaccard Coefficient Index (JC) and Dice Index (DI) [8].
The internal clustering index can select efficient clustering algorithm and optimal clustering number without additional information. In 1974, Calinski T and Harabasz J proposed the Calinski-Harabasz Index (CH), which evaluated the clustering effect by the tightness of the cluster and the tightness between the clusters [9]. In 1979, Davies D L and Bouldin D W proposed the Davies-Bouldin Index (DBI) to find the best clustering effect by calculating the ratio of the sum of the average distances between the two clusters and the distance between the cluster centers [10]. In 1986, Peter J. Rousseeuw proposed the Silhouette Coefficient, which combines the factors of intra-cluster polymerization and inter-cluster resolution to evaluate the clustering effect [11]. Dunn’s Index (DI) is an index based on cluster element distance proposed by J. Dunn in 1974. It is evaluated by calculating the ratio of the shortest distance of elements in any two different clusters to the maximum distance in any cluster [12]. However, DI evaluates the data of the ring distribution very poorly. Other internal clustering evaluation indexes include Root-mean-square std dev (RMSSTD) [13], R-squared (RS) [14], Xie-Beni index (XB) and other indexes [15]. The key to these indexes is the degree of separation, which represents whether the distance from the sample to the cluster is far enough. In this article I will use the abbreviations for these indexes.

2. Internal evaluation index

In this section I will study the Silhouette Index and the Calinski-Harabasz Index, which are the basis of the reference, and describe the algorithm steps and formulas for the proposed Peak Weight Index.

2.1. Silhouette index

The Silhouette Index obtains the optimal clustering number by the difference between the average distance within the cluster and the minimum distance between the clusters, that is, the optimal clustering effect, which is defined as follows.

\[
S = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{b(i) - a(i)}{\max \{a(i), b(i)\}} \right)
\]

among them: \(a(i)\) represents the average distance of sample \(i\) to other samples in the cluster, \(b(i)\) represents the minimum distance of the sample from \(i\) to the other clusters.

2.2. Calinski-Harabasz index

The Calinski-Harabasz Index is an evaluation index based on the degree of dispersion between clusters and clusters, and is defined as follows.

\[
CH(K) = \frac{B(K)}{W(K)} = \frac{\left( \sum_{k=1}^{K} a_k \| \bar{x}_k - \bar{x} \| \right)^2}{\sum_{c=1}^{K} \sum_{j=1}^{N} \left( x_j - \bar{x}_c \right)^2}
\]

among them: \(K\) is the corresponding number of clusters, \(B(K)\) is the inter-cluster divergence, also called the inter-cluster covariance, \(W(K)\) is the intra-cluster divergence, also called the intra-cluster covariance, and \(N\) is the number of samples. The larger the \(B(K)\) is, the higher the degree of dispersion between clusters is. The smaller the \(W(K)\) is, the closer the relationship is in the cluster. The higher the ratio is, the larger the value of the CH index is, that is, the better the clustering effect is.

2.3. Peak Weight index

The Peak Weight Index is an improved index based on the data separation and aggregation of the S and CH index. It solves the fluctuation of the index value caused by the sensitivity of the S and CH index to the original data, that is, to generate multiple optimal clustering numbers for the same clustering function [16]. Suppose there are three adjacent clusters A, B and C in one data set. Without considering the influence of other clusters, if they are clustered into \{A, B, C\} or \{A\}, \{B\} and \{C\}, the numerical performance of clustering indexes will be excellent. However, clustering into \{A, B\} or \{B, C\} will cause the clustering index to fluctuate, that is, the influence of data dispersion between clusters and clusters. The algorithm steps of PWI are as follows.

Step 1: Initialize the list of internal metrics applied to the clustering effect evaluation, and set the initial weight 1 for the evaluation points that may appear in each metric.
Step 2: The peak points in the range coefficient $\varepsilon$ are traversed from $i=2$, and the corresponding peak points are weighted according to the three principles.

Step 3: Calculate the corresponding weight calculation formula for each index.

Step 4: Take the mean value of $S$ and $CH$ as the value of $PWI$, and select the optimal cluster number or the best clustering effect according to the standard. The definition of $PWI$ is as follows.

\[
PWS(K) = \frac{\sum_{i=K-\varepsilon}^{K+\varepsilon} x_i p_i}{n}
\]

\[
PWC(K) = \frac{\sum_{i=K-\varepsilon}^{K+\varepsilon} x_i p_i}{m}
\]

\[
PWI(K) = \frac{PWS(K) + PWC(K)}{2}
\]

Among them: $K$ represents the number of corresponding clusters, and $\varepsilon$ represents the peak selection range, which generally depends on the collected data set. In the following experiment, the default value is 1, and $p_i$ represents the weight of the corresponding index point. In the index, the boundary of the index point where the peak appears is also introduced. The upper and lower limits are generally selected as $2\sigma$ and $2\sigma^2$, respectively, but in practice, the boundary should be appropriately selected according to the distribution of the data set. In the experiment, we used the maximum and minimum normalization for the value of $CH$, and then calculated [17].

Three principles corresponding to $PWI$: 1. If the predecessor and successor nodes of the evaluated index points are not peak nodes, and the increase or decrease between nodes shows a negative correlation, it will not be included in the evaluation of weights. 2. The difference between the index point of the evaluation and the front and rear nodes is less than the lower limit, and the weight is not considered. 3. The difference between the index point of the evaluation and the front and back nodes is greater than the upper limit, and the node does not consider it.

3. Influencing factors and experiments

In this section, four possible factors that may affect the evaluation effect are proposed: monotonic, outlier, separability and proximity. At the same time, we used three indexes to carry out simulation experiments on four self-built corresponding data sets, the possible influences of four factors on the indexes were studied, and to conduct further research on the two real data sets.

3.1. Monotonic factor

In this experiment, data sets with a single factor are used, and a k-means clustering algorithm is used to separate good data sets and obtain different cluster numbers [18]. As shown in Figure 1, the data set consists of three nearly perfect clusters. The internal connections of the clusters are relatively tight, and the connections between clusters are relatively discrete. From Table 1, we can see that the optimal cluster number of the three indexes for the data set is 3, and the effect is good. It can be known that monotonicity is not the influencing factor affecting the evaluation of clustering effect.

Table 1. The effect of monotonicity

|    | S   | CH  | PWI |
|----|-----|-----|-----|
| 2  | 0.641 | 154 | 0.541 |
| 3  | **0.883** | **2725** | **0.819** |
| 4  | 0.728 | 2150 | 0.783 |
| 5  | 0.604 | 1977 | 0.662 |
| 6  | 0.428 | 2021 | 0.610 |
| 7  | 0.443 | 2072 | 0.592 |
| 8  | 0.452 | 2099 | 0.602 |

Figure 1. Monotone data set

3.2. Outlier factor

In order to study the influence of outliers on clustering indexes, we added a small number of outlier nodes based on the data set of monotonic factors [19], as shown in Figure 2.
From Table 2, it can be found that CH considers that the optimal cluster number is 4, which is affected by the outliers, and the optimal cluster number of S and PWI is 3, which is better. Therefore, we believe that outliers are one of the factors affecting clustering indexes.

### Table 2. The effect of outlier

|   | S      | CH    | PWI   |
|---|--------|-------|-------|
| 2 | 0.632  | 163   | 0.525 |
| 3 | 0.836  | 1369  | 0.851 |
| 4 | 0.770  | 1488  | 0.829 |
| 5 | 0.626  | 1264  | 0.770 |
| 6 | 0.604  | 1211  | 0.704 |
| 7 | 0.572  | 1220  | 0.673 |
| 8 | 0.424  | 1288  | 0.660 |

Figure 2. Outlier data set

#### 3.3. Separability factor

Discreteness is mainly to study the influence of the tightness of nodes in the cluster on the evaluation of clustering effect. We constructed the discrete dataset as shown in Figure 3. It can be seen from the figure that the internal nodes of the three clusters are relatively dispersed.

According to the experimental results in Table 3, the optimal clustering number of S, CH and PWI is 3, that is, the influence of discreteness on the three clustering evaluation indexes is not obvious.

### Table 3. The effect of separability

|   | S      | CH    | PWI   |
|---|--------|-------|-------|
| 2 | 0.647  | 235   | 0.508 |
| 3 | 0.738  | 659   | 0.808 |
| 4 | 0.614  | 616   | 0.731 |
| 5 | 0.462  | 521   | 0.584 |
| 6 | 0.315  | 465   | 0.514 |
| 7 | 0.309  | 433   | 0.382 |
| 8 | 0.282  | 397   | 0.333 |

Figure 3. Separability data set

#### 3.4. Proximity factor

Proximity is the study of the degree of separation between clusters and clusters [20]. As shown in Figure 4, we can see that there are two closely clustered clusters, and the proximity considers the impact of close association between clusters on the evaluation of clustering indexes.

### Table 4. The effect of proximity

|   | S     | CH    | PWI   |
|---|-------|-------|-------|
| 2 | 0.727 | 443   | 0.412 |
| 3 | 0.724 | 871   | 0.731 |
| 4 | 0.573 | 686   | 0.579 |
| 5 | 0.400 | 525   | 0.367 |
| 6 | 0.291 | 506   | 0.250 |
| 7 | 0.298 | 501   | 0.192 |
| 8 | 0.255 | 436   | 0.181 |

Figure 4. Proximity data set

From Table 4, we can find that the optimal cluster number of S index is 2, but it is similar to the number of cluster clusters of 3, while CH and PWI successfully choose the best cluster number and
perform well. It can be seen that the S index is sensitive to proximity, and proximity is one of the factors affecting the evaluation of clustering effects.

3.5. Experiment with real data sets
The two real data sets used in this experiment are the classic dataset from UCI, Iris, and a map of the Chinese community in Shanghai, which is crawled from the Internet. Iris data set has 150 objects, including sepal length, sepal width, petal length and petal width. First, I used Principal Component Analysis (PCA) to reduce the dimensionality of these objects. The data distribution after dimensionality reduction is shown in Figure 5.

Table 5. The evaluation results of Iris

|   | S   | CH  | PWI  |
|---|-----|-----|------|
| 2 | 0.711 | 557 | 0.720 |
| 3 | 0.494 | 452 | 0.622 |
| 4 | 0.378 | 387 | 0.423 |
| 5 | 0.289 | 307 | 0.301 |
| 6 | 0.260 | 289 | 0.217 |
| 7 | 0.248 | 266 | 0.160 |
| 8 | 0.174 | 233 | 0.120 |

Figure 5. The original distribution of Iris

We can see from the figure that the reduced-dimensional data is mainly clustered into two clusters. According to the external label information attached to the dataset, we can speculate that there are two clusters in the original data with node crossover, that is, we default the data distribution after dimensionality reduction to two clusters, which contain many influencing factors. According to the experimental results in Table 5, the S, CH and PWI can select the best cluster number and perform well.

For the second real data set, coordinate data of 350 communities from five regions in Shanghai, China are collected, and the data distribution is shown in Figure 6. In the figure, we can divide the data into five clusters by visual aid, and the five clusters are complicated with each other and are affected by many factors. From the experimental results in Table 6, we can find that the S index clusters the data into two clusters, and fluctuates at k=4 and k=6, the evaluation effect is not good. However, CH is to cluster the data into six clusters. Although it is close to the real result, the optimal cluster number is not selected. Only by combining the characteristics of the two indicators, PWI adopts the method of assigning weights to the places where there are fluctuations in the evaluation process, and the optimal cluster number is 5.

Table 6. The evaluation results of coordinates

|   | S   | CH  | PWI  |
|---|-----|-----|------|
| 2 | 0.872 | 6337 | 0.445 |
| 3 | 0.750 | 3764 | 0.619 |
| 4 | 0.819 | 19232 | 0.439 |
| 5 | 0.722 | 14605 | **0.905** |
| 6 | 0.823 | **80578** | 0.723 |
| 7 | 0.757 | 72578 | 0.837 |
| 8 | 0.710 | 68165 | 0.774 |

Figure 6. The original distribution of coordinates

4. Concluding remarks
As an unsupervised learning method, clustering requires no training or extra labels, so it is highly flexible, but it is also more difficult to evaluate the clustering effect. This paper mainly studies the internal clustering effect evaluation, and proposes an improved index PWI based on the internal
clustering evaluation of S index and CH index. The factors that may affect the evaluation effect are studied from the aspects of monotonicity, outliers, separability and proximity. Then, through experiments on four corresponding self-created data sets and two real data sets, the evaluation performance of the internal clustering evaluation index is studied, which proves that PWI has a good evaluation effect on the six data sets. However, further research and experiments are needed for some details such as the selection range of ε in PWI.

References
[1] Clausi D A. K-means Iterative Fisher (KIF) unsupervised clustering algorithm applied to image texture segmentation[J]. Pattern Recognition, 2002, 35(9):1959-1972.
[2] Benítez I, Quijano A, Díez J L, et al. Dynamic clustering segmentation applied to load profiles of energy consumption from Spanish customers[J]. International Journal of Electrical Power & Energy Systems, 2014, 55(2):437-448.
[3] Neely W B, Branson D R, Blau G E. Partition coefficient to measure bioconcentration potential of organic chemicals in fish[J]. Environmental Science & Technology, 1974, 8(13):1113-1115.
[4] Campello R J G B. A fuzzy extension of the Rand index and other related indexes for clustering and classification assessment[J]. Pattern Recognition Letters, 2007, 28(7):833-841.
[5] Santos J M, Embrechts M. On the Use of the Adjusted Rand Index as a Metric for Evaluating Supervised Classification[C]// Artificial Neural Networks-icann, International Conference, Limassol, Cyprus, September. 2009.
[6] Maes F, Collignon A, Vandermeulen D, et al. Multi-Modality Image Registration Maximization of Mutual Information[J]. Proceedings of Mmbia, 1996, 16(2):14-22.
[7] A. F. L. Nemec and , R. O. Brinkhurst. The Fowlkes–Mallows Statistic and the Comparison of Two Independently Determined Dendrograms[J]. Canadian Journal of Fisheries & Aquatic Sciences, 1988, 45(6):971-975.
[8] Kogge P M. Jaccard Coefficients as a Potential Graph Benchmark.[C]// IEEE International Parallel & Distributed Processing Symposium Workshops. 2016.
[9] Calinski T , Harabasz J . A dendrite method for cluster analysis[J]. Communications in Statistics, 1974, 3(1):1-27.
[10] Davies D L , Bouldin D W . A Cluster Separation Measure[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1979, PAMI-1(2):224-227.
[11] Rousseeuw P J . Silhouettes : a graphical aid to the interpretation and validation of cluster analysis[J]. J. Comput. Appl. Math. 1987, 20.
[12] J. Dunn, “Well separated clusters and optimal fuzzy partitions,” J. Cybern., vol. 4, no. 1, pp. 95–104, 1974.
[13] MarkMarcucci. Applied Multivariate Techniques[J]. Technometrics, 1996, 39(1):100-101.
[14] Halkidi M, Batistakis Y, Vazirgiannis M. On Clustering Validation Techniques[J]. Journal of Intelligent Information Systems Integrating Artificial Intelligence & Database Technologies, 2001, 17(2-3):107-145.
[15] Xie X L , Beni G . A validity measure for fuzzy clustering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1991, 13(8):841-847.
[16] Wang X, Wang Y, Wang L. Improving fuzzy -means clustering based on feature-weight learning[J]. Pattern Recognition Letters, 2004, 25(10):1123-1132.
[17] Srivastava M S, Keen K J. Estimation of the Interclass Correlation Coefficient[J]. Annals of Human Genetics, 2012, 57(2):159-165.
[18] Błaszczynski J, Greco S, Słowiński R, et al. Monotonic Variable Consistency Rough Set Approaches[J]. International Journal of Approximate Reasoning, 2009, 50(7):979-999.
[19] Aggarwal C C, Yu P S. Outlier detection for high dimensional data[J]. Proc Acm Sigmod May Santa Barbara, 2001, 30(2):37-46.
[20] Krauthgamer R, Lee J R. Navigating nets:simple algorithms for proximity search[C]// Fifteenth Acm-siam Symposium on Discrete Algorithms. 2004.