A Clustering Validity Assessment Index

Youngok Kim and Soowon Lee

School of Computing, Soongsil University
1-1 Sang-Do Dong, Dong-Jak Gu, Seoul 156-743 Korea
{yokim, swlee}@computing.ssu.ac.kr
http://mining.ssu.ac.kr/

Abstract. Clustering is a method for grouping objects with similar patterns and finding meaningful clusters in a data set. There exist a large number of clustering algorithms in the literature, and the results of clustering even in a particular algorithm vary according to its input parameters such as the number of clusters, field weights, similarity measures, the number of passes, etc. Thus, it is important to effectively evaluate the clustering results a priori, so that the generated clusters are more close to the real partition. In this paper, an improved clustering validity assessment index is proposed based on a new density function for inter-cluster similarity and a new scatter function for intra-cluster similarity. Experimental results show the effectiveness of the proposed index on the data sets under consideration regardless of the choice of a clustering algorithm.

1 Introduction

Clustering is a method for grouping objects with similar patterns and finding meaningful clusters in a data set. Clustering is important in many fields including data mining, information retrieval, pattern recognition, etc [1,2,3]. A wide variety of algorithms have been proposed in the literature for different applications and types of data set [1,2,4]. These algorithms can be classified according to the clustering methods used, such as partitional vs. hierarchical, monothetic vs. polythetic, and so on [1]. However, even a particular clustering algorithm behave in a different way depending on its features of the data set, field weights, similarity measure, etc [1,2,4]. A more critical issue is to decide the optimal number of clusters that best fits the underlying data set. Thus, it is important to effectively evaluate the clustering results a priori, so that the generated clusters are more close to the real partition.

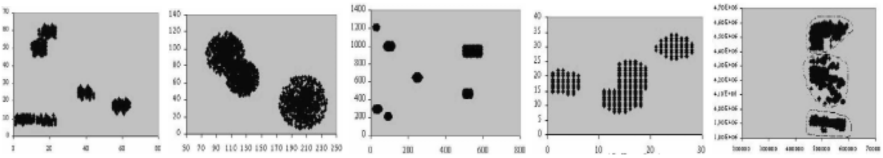
The goal of this paper is to propose a new clustering validity index, *S_Dbw**, which takes advantage of the concept underlying *S_Dbw* index [3], while resolving problems in *S_Dbw* index. *S_Dbw** index can deal with inter-cluster similarity and intra-cluster similarity more robustly and enables the selection of optimal number of clusters more effectively for a clustering algorithm. The rest of this paper is organized as follows. Section 2 presents the related work and Section 3 defines a new validity index, *S_Dbw**. Then, in Section 4, we show the experimental results of *S_Dbw** index com

K.-Y. Whang, J. Jeon, K. Shim, J. Srivatava (Eds.): PAKDD 2003, LNAI 2637, pp. 602–608, 2003.

pared with the results of *S_Dbw* index. Finally, in Section 5, we conclude by briefly presenting our contributions and provide directions for further research.

2 Related Work

There exist many indices for clustering validity assessment including *Dunn and Dunn-like* indices [5], *DB*(the Davies-Bouldin) index [6], *RMSSDT*(Root-mean-square standard deviation of the new cluster) index [7], *SPR*(Semi-partial R-squared) index [7], *RS*(R-squared) index [7], *CD*(Distance between two clusters) index [7], *S_Dbw* (Scatter and Density between clusters) index [3], etc. Recently research shows that *S_Dbw* index has better performance and effectiveness than other indices [2,3].



(a) DataSet1 (b) DataSet2 (c) DataSet3 (d) DataSet4 (e) RealData1

Fig. 1. Sample synthetic datasets (a,b,c,d) and a real dataset(e) [3].

Fig. 1 depicts five data sets used in [3]. Intuitively, DataSet1 has four clusters, DataSet2 has two or three clusters, DataSet3 has seven clusters, DataSet4 has three clusters, and RealData1 has three clusters, respectively. Table 1 shows the optimal number of clusters proposed by *RS*, *RMSSTD*, *DB*, *SD*, *S_Dbw* indices for the data sets given in Fig. 1, indicating that *S_Dbw* index can find optimal clusters for all given data sets [3].

Table 1. Optimal number of clusters proposed by different validity indices [3].

Index	**DataSet1**	**DataSet2**	**Dataset3**	**DataSet4**	**RealData1**
			Optimal number of clusters		
RS, RMSSTD	3	2	5	4	3
DB	6	3	7	4	3
SD	4	3	6	3	3
S_Dbw	4	2	7	3	3

3 *S_Dbw** Index

The definition of *S_Dbw* index indicates that both criteria of inter-cluster and intra-cluster similarity are properly combined, enabling reliable evaluation of clustering results [3]. Though the definition of *S_Dbw* index is proper for both criteria of inter-cluster and intra-cluster similarity, *S_Dbw* index has some problems. First, *Dens_bw* of *S_Dbw* index for inter-cluster similarity finds the neighborhood simply within the boundary of *stdev*. Thus, the detected neighborhood does not reflect the inter-cluster similarity precisely in the case of non-circular clusters. Second, *Scat* of *S_Dbw* index for intra-cluster similarity measures the scattering by calculating relative ratio between the average variance of clusters and the variance of the entire data set. The problem here is *Scat* does not consider the number of tuples in each cluster. If two clusters' variances are almost equal, their variances have the same effect on *Scat* value, regardless of the number of tuples in the clusters. Moreover, since *Scat* value is calculated by averaging the variances of clusters, *S_Dbw* tends to prefer small divided clustering to large combined clustering. These properties of *S_Dbw* index may effect negatively on finding good partition.

To overcome the problems of *S_Dbw*, a new index S_Dbw^* is defined in this paper. S_Dbw^* is based on the same clustering evaluation concepts as *S_Dbw* (inter-, intra-cluster similarity). However, S_Dbw^* index is different from *S_Dbw* index in the sense that it finds the neighborhood by using the confidence interval of each dimension rather than a hyper-sphere, so that the size and the shape of each cluster is reflected in the density function. Another difference is that S_Dbw^* index measures the weighted average of scattering within clusters, so that the data ratio of each cluster is reflected in the scatter function.

The inter-cluster similarity of S_Dbw^* index is defined as follows. Let *S* be a data set, and *n* be the number of tuples in *S*. For a given data point *m*, the term $density^*(m)$ is defined similarly as in equation (1):

$$density^*(m) = \sum_{l=1}^{n} f^*(x_l, m) \ . \tag{1}$$

Here, the function $f^*(x,m)$ denotes whether the distance between two data points *x* and *m* is less or equal to the confidence interval for each dimension. Let *k* be the number of dimensions for the data set, then $f^*(x,m)$ is defined as:

$$f^*(x,m) = \begin{cases} 1 : CI^p \leq d(x^p,m^p) \leq CI^p, \ (\forall p,\ 1 \leq p \leq k) \\ 0 : otherwise \end{cases} \tag{2}$$

where CI^p denotes the confidence intervals of p-th dimension. With the confidence interval of 95%, CI^p is defined as follows:

$$CI^p = u^p \pm (1.96 \times \frac{\sigma^p}{\sqrt{n}}) \tag{3}$$

where u^p and σ^p represent the average and the standard deviation of p-th dimension, respectively. Let m_{ij} be the middle point of the line segment defined by the centers of clusters c_i, c_j. Then, terms in equation (3) are substituted with u_{ij}^p, σ_{ij}^p, n_{ij}, where

$$u_{ij}^p = \frac{u_i^p + u_j^p}{2}, \sigma_{ij}^p = \frac{\sigma_i^p + \sigma_j^p}{2}, n_{ij} = n_i + n_j \ . \tag{4}$$

From the above definitions, $Dens\_bw^*(c)$ is defined as in equation (5):

$$Dens\_bw^*(c) = \frac{1}{c(c-1)} \sum_{i=1}^{c} \left[\sum_{\substack{j=1 \\ i \neq j}}^{c} \frac{density^*(m_{ij})}{\max\{density^*(v_i), density^*(v_j)\}} \right] . \tag{5}$$

The basic idea of *Dens_bw** is that it detects the neighborhood by using the confidence interval for each dimension instead of *stdev* boundary, resulting the different boundary for each dimension according to the cluster's shape. This implies that *Dens_bw** is more flexible than *Dens_bw* of *S_Dbw* index in detecting neighborhood in case of non-circular or lined up clusters.

For the intra-cluster similarity of *S_Dbw** index, *Scat***(c)* is defined as:

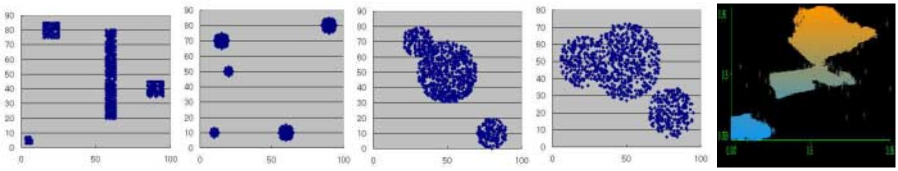
$$Scat^*(c) = \frac{1}{c} \sum_{i=1}^{c} \frac{n-n_i}{n} \left\| \sigma(v_i)^2 \right\| \Big/ \left\| \sigma(S)^2 \right\| \ , \quad where \ \left\| x \right\| = (xx^T)^{1/2} \tag{6}$$

where n is the total number of tuples in S and n_i is the number of tuples in cluster c_i.

If two clusters have the same variance but different numbers of tuples, the variance of cluster with more tuples has more effect on the scatter function than the one with less tuples. By minimizing the influence of small clusters or noisy clusters, *Scat** can be more robust than *Scat*.of *S_Dbw* index.

From equation (5) and (6), we have *S_Dbw***(c)* index, which is defined as follows:

$$S\_Dbw^*(c) = Dens\_bw^*(c) + Scat^*(c) \ . \tag{7}$$



(a)DataSet1 (b)DataSet2 (c)DataSet3 (d)DataSet4 (e)RealData

Fig. 2. Experimental data sets((a)-(d)) and a real data set(e).

4 Experiment

In this section, we experimentally evaluate the proposed validity index, *S_Dbw**, in comparison with *S_Dbw* index. Fig. 2 shows the experimental data sets consisting of four artificial data sets and one real data set. Intuitively, DataSet1 has four clusters, but one cluster consists of lined up points. DataSet2 has five circular clusters. DataSet3 has three clusters, but two clusters' boundaries are very closely connected. DataSet4 has roughly two clusters, but all points spread broadly. RealData is the North East data set which has 50,000 postal addresses(points) representing three metropolitan areas(New York, Philadelphia and Boston), hence three clusters are in the form of uniformly distributed rural areas and smaller population centers [8].

S_Dbw[*] index is evaluated in comparison with *S_Dbw* index experimentally by applying three representative clustering algorithms from different categories. Clustering algorithms used in the experiments are K-Means(partition-based), EM(statistical model-based), and SOM(Kohonen network-based). For each algorithm, the given number of clusters varies from two to ten. We applied EM and SOM algorithms to DataSet1 and DataSet4, since K-Means algorithm is not appropriate in this case because of its weakness on non-circular and sparse data sets. EM and K-Means algorithms are applied to DataSet2 and DataSet3 to have circular clusters. RealData is evaluated by EM algorithm.

Table 2. Optimal number of clusters found by *S_Dbw* and *S_Dbw*[*] for each algorithm and data.

Data	Algo-rithm	Index	Number of Clusters								
			Optimal								Not Optimal
Data Set1	EM	*S_Dbw*	4	5	8	6	7	9	10	3	2
			0.125	0.155	0.167	0.178	0.181	0.240	0.312	0.503	0.740
		S_Dbw[*]	4	5	6	8	3	2	10	7	9
			0.089	0.109	0.111	0.364	0.409	0.455	0.512	0.520	0.538
	SOM	*S_Dbw*	10	6	7	9	4	5	8	3	2
			0.133	0.148	0.160	0.163	0.167	0.168	0.175	0.597	0.740
		S_Dbw[*]	4	5	8	7	10	9	6	3	2
			0.123	0.228	0.238	0.252	0.307	0.326	0.370	0.428	0.455
Data Set2	EM	*S_Dbw*	8	4	6	7	9	3	5	10	2
			0.003	0.019	0.134	0.135	0.139	0.145	0.161	0.380	0.716
		S_Dbw[*]	8	9	5	7	4	10	6	3	2
			0.003	0.122	0.156	0.182	0.273	0.292	0.301	0.331	2.353
	K-Means	*S_Dbw*	5	4	6	2	7	8	9	10	3
			0.004	0.019	0.077	0.087	0.099	0.112	0.138	0.142	0.149
		S_Dbw[*]	5	6	9	7	8	10	4	3	2
			0.080	0.019	0.077	0.087	0.099	0.112	0.138	0.142	0.149
Data Set3	EM	*S_Dbw*	10	9	8	6	7	4	5	2	3
			0.259	0.281	0.317	0.358	0.376	0.409	0.413	0.473	0.514
		S_Dbw[*]	3	4	10	6	2	9	8	7	5
			0.272	0.306	0.384	0.400	0.442	0.473	0.488	0.509	0.556
	K-Means	*S_Dbw*	10	9	5	6	2	3	4	8	7
			0.306	0.310	0.440	0.472	0.473	0.482	0.491	0.505	0.530
		S_Dbw[*]	3	9	2	4	8	7	10	6	5
			0.281	0.436	0.442	0.442	0.514	0.514	0.550	0.553	0.558
Data Set4	EM	*S_Dbw*	10	9	8	6	7	5	4	3	2
			0.309	0.317	0.355	0.398	0.420	0.429	0.528	0.722	1.02
		S_Dbw[*]	9	2	10	8	5	7	3	6	4
			0.503	0.558	0.561	0.598	0.611	0.648	0.659	0.703	0.907
	SOM	*S_Dbw*	9	10	8	6	7	5	4	3	2
			0.328	0.329	0.350	0.390	0.445	0.608	0.703	0.934	0.989
		S_Dbw[*]	2	9	8	10	5	7	4	3	6
			0.324	0.482	0.487	0.500	0.616	0.670	0.683	0.707	0.712
Real Data	EM	*S_Dbw*	4	7	9	2	8	10	5	6	3
			1.21	1.31	1.33	1.34	1.34	1.35	1.57	1.62	1.66
		S_Dbw[*]	3	4	9	8	5	2	7	10	6
			0.250	0.296	0.348	0.353	0.406	0.421	0.422	0.459	0.468

Table 2 compares the ranks and index values evaluated by *S_Dbw* and *S_Dbw**. For DataSet1, 3, 4 and RealData, two indices show the remarkable different results. For Dataset1, *S_Dbw* index evaluates ten clusters as the best one in SOM, while *S_Dbw** index evaluates four clusters as the best one. For DataSet3, *S_Dbw* index evaluates ten clusters as the best one both in EM and K-Means (and evaluates three clusters as the worst one in EM and the sixth one in SOM), while *S_Dbw** evaluated three clusters as the best one in both algorithms. For Dataset4, *S_Dbw* index evaluated nine clusters as the best one and two clusters as the worst one in SOM, while *S_Dbw** finds two clusters correctly. Also, for RealData, *S_Dbw* index evaluated four clusters as the best one, while *S_Dbw** finds three clusters exactly.

In DataSet2, EM algorithm finds five clusters when the number of clusters is given as eight. EM is known to be an appropriate optimization algorithm for constructing proper statistical models on the data. It assigns the density probability of each data to be included in a cluster according to the constructed model. Though EM algorithm calculates the density probability for eight clusters, data points are assigned to only five clusters in this case.

5 Conclusions

In this paper we addressed the important issue of assessing the validity of clustering algorithms' results. We defined a new validity index, *S_Dbw**, for assessing the results of clustering algorithms. For inter-cluster similarity, *S_Dbw** uses the confidence interval for each dimension and for each cluster, while for intra-cluster similarity, *S_Dbw** uses weighted average variance between clusters. We evaluated the flexibility of *S_Dbw** index experimentally on five data sets. The results implies that *S_Dbw** index outperforms *S_Dbw* index for all clustering algorithm used in the experiments, especially when the real partition is noncircular or sparse.

The essence of clustering is not a totally resolved issue, and depending on the application domain, we may consider different aspects as more significant ones. Having an indication that a good partitioning can be proposed by the cluster validity index, the domain experts may analyze further the validation procedure.

References

1. Jain, A. K., Murty, M.N., Flynn, P.J.: Data Clustering: A Review. *ACM Computing Surveys*, Vol. 31. No. 3 (1999) 264–323
2. Halkidi, M., Batistakis, I., Varzirgiannis, M.: On Clustering Validation Techniques, *Journal of Intelligent Information Systems*, Vol. 17. No. 2-3 (2001) 107–145
3. Halkidi, M., Varzirgiannis, M.: Clustering Validity Assesment: Finding the Optimal Partitioning of a Data Set. *ICDM* (2001) 187–194
4. Fasulo, D.: An Analysis of Recent Work on Clustering Algorithms. *Technical Report*, University of Washington (1999)

5. Dunn, J. C.: Well Separated Clusters and Optimal Fuzzy Partitions. J. Cybern. Vol. 4 (1974) 95–104
6. Davies, DL., Douldin, D. W.: A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Vol. 1. No. 2 (1979)
7. Sharma, S.C.: Applied Multivariate Techniques. John Willy & Sons (1996)
8. Theodoridis, Y.: Spatial Datasets – an unofficial collection. (1996) http://dias.cti.gr/~ytheod/research/datasets/spatial.html