Understanding of Internal Clustering Validation Measures

Yanchi Liu[1,2], Zhongmou Li[2], Hui Xiong[2], Xuedong Gao[1], Junjie Wu[3]
[1]*School of Economics and Management, University of Science and Technology Beijing, China*
liuyanchi@manage.ustb.edu.cn, gaoxuedong@manage.ustb.edu.cn
[2]*MSIS Department, Rutgers Business School, Rutgers University, USA*
mosesli@pegasus.rutgers.edu, hxiong@rutgers.edu
[3]*School of Economics and Management, Beihang University, China*
wujj@buaa.edu.cn

***Abstract*—Clustering validation has long been recognized as one of the vital issues essential to the success of clustering applications. In general, clustering validation can be categorized into two classes, external clustering validation and internal clustering validation. In this paper, we focus on internal clustering validation and present a detailed study of 11 widely used internal clustering validation measures for crisp clustering. From five conventional aspects of clustering, we investigate their validation properties. Experiment results show that S_Dbw is the only internal validation measure which performs well in all five aspects, while other measures have certain limitations in different application scenarios.**

I. Introduction

Clustering, one of the most important unsupervised learning problems, is the task of dividing a set of objects into clusters such that objects within the same cluster are similar while objects in different clusters are distinct. Clustering is widely used in many fields, such as image analysis and bioinformatics. As an unsupervised learning task, it is necessary to find a way to validate the goodness of partitions after clustering. Otherwise, it would be difficult to make use of different clustering results.

Clustering validation, which evaluates the goodness of clustering results [1], has long been recognized as one of the vital issues essential to the success of clustering applications [2]. External clustering validation and internal clustering validation are the two main categories of clustering validation. The main difference is whether or not external information is used for clustering validation. An example of external validation measure is entropy, which evaluates the "purity" of clusters based on the given class labels [3].

Unlike external validation measures, which use external information not present in the data, internal validation measures only rely on information in the data. The internal measures evaluate the goodness of a clustering structure without respect to external information [4]. Since external validation measures know the "true" cluster number in advance, they are mainly used for choosing an optimal clustering algorithm on a specific data set. On the other hand, internal validation measures can be used to choose the best clustering algorithm as well as the optimal cluster number without any additional information. In practice, external information such as class labels is often not available in many application scenarios. Therefore, in the situation that there is no external information available, internal validation measures are the only option for cluster validation.

In literature, a number of internal clustering validation measures for crisp clustering have been proposed, such as CH, I, DB, SD and S_Dbw. However, current existing measures can be affected by various data characteristics. For example, noise in data can have a significant impact on the performance of an internal validation measure, if minimum or maximum pairwise distances are used in the measure. The performance of existing measures in different situations remains unknown. Therefore, we present a detailed study of 11 widely used internal validation measures, as shown in Table I. We investigate their validation properties in five different aspects: monotonicity, noise, density, subclusters and skewed distributions. For each aspect, we generate synthetic data for experiments. These synthetic data well represent the properties. Finally, the experiment results show that S_Dbw is the only internal validation measure which performs well in all five aspects, while other measures have certain limitations in different application scenarios, mainly in aspects of noise and subclusters.

II. Internal Clustering Validation Measures

In this section, we introduce some basic concepts of internal validation measures, as well as a suite of 11 widely used internal validation indices.

As the goal of clustering is to make objects within the same cluster similar and objects in different clusters distinct, internal validation measures are often based on the following two criteria [4] [5].

I. Compactness. It measures how closely related the objects in a cluster are. A group of measures evaluate cluster compactness based on variance. Lower variance indicates better compactness. Also, there are numerous measures estimate the cluster compactness based on distance, such as maximum or average pairwise distance, and maximum or average center-based distance.

1550-4786/10 $26.00 © 2010 IEEE
DOI 10.1109/ICDM.2010.35

Table I
INTERNAL CLUSTERING VALIDATION MEASURES

	Measure	Notation	Definition	Optimal value
1	Root-mean-square std dev	$RMSSTD$	$\{\sum_i \sum_{x \in C_i} \Vert x - c_i \Vert^2 / [P \sum_i (n_i - 1)]\}^{\frac{1}{2}}$	Elbow
2	R-squared	RS	$(\sum_{x \in D} \Vert x - c \Vert^2 - \sum_i \sum_{x \in C_i} \Vert x - c_i \Vert^2) / \sum_{x \in D} \Vert x - c \Vert^2$	Elbow
3	Modified Hubert Γ statistic	Γ	$\frac{2}{n(n-1)} \sum_{x \in D} \sum_{y \in D} d(x,y) d_{x \in C_i, y \in C_j}(c_i, c_j)$	Elbow
4	Calinski-Harabasz index	CH	$\frac{\sum_i n_i d^2(c_i,c)/(NC-1)}{\sum_i \sum_{x \in C_i} d^2(x,c_i)/(n-NC)}$	Max
5	I index	I	$(\frac{1}{NC} \cdot \frac{\sum_{x \in D} d(x,c)}{\sum_i \sum_{x \in C_i} d(x,c_i)} \cdot \max_{i,j} d(c_i, c_j))^p$	Max
6	Dunn's indices	D	$\min_i \{\min_j (\frac{\min_{x \in C_i, y \in C_j} d(x,y)}{\max_k \{\max_{x,y \in C_k} d(x,y)\}})\}$	Max
7	Silhouette index	S	$\frac{1}{NC} \sum_i \{\frac{1}{n_i} \sum_{x \in C_i} \frac{b(x)-a(x)}{\max[b(x),a(x)]}\}$	Max
			$a(x) = \frac{1}{n_i-1} \sum_{y \in C_i, y \neq x} d(x,y), b(x) = \min_{j,j \neq i} [\frac{1}{n_j} \sum_{y \in C_j} d(x,y)]$	
8	Davies-Bouldin index	DB	$\frac{1}{NC} \sum_i \max_{j,j \neq i} \{[\frac{1}{n_i} \sum_{x \in C_i} d(x,c_i) + \frac{1}{n_j} \sum_{x \in C_j} d(x,c_j)] / d(c_i,c_j)\}$	Min
9	Xie-Beni index	XB	$[\sum_i \sum_{x \in C_i} d^2(x,c_i)] / [n \cdot min_{i,j \neq i} d^2(c_i,c_j)]$	Min
10	SD validity index	SD	$Dis(NC_{max}) Scat(NC) + Dis(NC)$	Min
			$Scat(NC) = \frac{1}{NC} \sum_i \Vert \sigma(C_i) \Vert / \Vert \sigma(D) \Vert, Dis(NC) = \frac{max_{i,j} d(c_i,c_j)}{min_{i,j} d(c_i,c_j)} \sum_i (\sum_j d(c_i,c_j))^{-1}$	
11	S_Dbw validity index	S_Dbw	$Scat(NC) + Dens\_bw(NC)$	Min
			$Dens\_bw(NC) = \frac{1}{NC(NC-1)} \sum_i [\sum_{j,j \neq i} \frac{\sum_{x \in C_i \bigcup C_j} f(x,u_{ij})}{max\{\sum_{x \in C_i} f(x,c_i), \sum_{x \in C_j} f(x,c_j)\}}]$	

D: data set; n: number of objects in D; c: center of D; P: attributes number of D; NC: number of clusters; C_i: the i–th cluster; n_i: number of objects in C_i; c_i: center of C_i; $\sigma(C_i)$: variance vector of C_i; $d(x,y)$: distance between x and y; $\Vert X_i \Vert = (X_i^T \cdot X_i)^{\frac{1}{2}}$

II. Separation. It measures how distinct or well-separated a cluster is from other clusters. For example, the pairwise distances between cluster centers or the pairwise minimum distances between objects in different clusters are widely used as measures of separation. Also, measures based on density are used in some indices.

The general procedure to determine the best partition and optimal cluster number of a set of objects by using internal validation measures is as follows.

Step 1: Initialize a list of clustering algorithms which will be applied to the data set.

Step 2: For each clustering algorithm, use different combinations of parameters to get different clustering results.

Step 3: Compute the corresponding internal validation index of each partition obtained in step 2.

Step 4: Choose the best partition and the optimal cluster number according to the criteria.

Table I shows a suite of 11 widely used internal validation measures. To the best of our knowledge, these measures represent a good coverage of the validation measures available in different fields, such as data mining, information retrieval, and machine learning. The "Definition" column gives the computation forms of the measures. Next, we briefly introduce these measures.

Most indices consider both of the evaluation criteria (compactness and separation) in the way of ratio or summation, such as DB, XB, and S_Dbw. On the other hand, some indices only consider one aspect, such as $RMSSTD$, RS, and Γ.

The Root-mean-square standard deviation ($RMSSTD$) is the square root of the pooled sample variance of all the attributes [6]. It measures the homogeneity of the formed clusters. R-squared (RS) is the ratio of sum of squares between clusters to the total sum of squares of the whole data set. It measures the degree of difference between clusters [6] [7]. The Modified Hubert Γ statistic (Γ) [8] evaluates the difference between clusters by counting the disagreements of pairs of data objects in two partitions.

The Calinski-Harabasz index (CH) [9] evaluates the cluster validity based on the average between- and within-cluster sum of squares. Index I (I) [1] measures separation based on the maximum distance between cluster centers, and measures compactness based on the sum of distances between objects and their cluster center. Dunn's index (D) [10] uses the minimum pairwise distance between objects in different clusters as the inter-cluster separation and the maximum diameter among all clusters as the intra-cluster compactness. These three indices take a form of $Index = (a \cdot Separation)/(b \cdot Compactness)$, where a and b are weights. The optimal cluster number is determined by maximizing the value of these indices.

The Silhouette index (S) [11] validates the clustering performance based on the pairwise difference of between- and within-cluster distances. In addition, the optimal cluster number is determined by maximizing the value of this index.

The Davies-Bouldin index (DB) [12] is calculated as follows. For each cluster C, the similarities between C and all other clusters are computed, and the highest value is assigned to C as its cluster similarity. Then the DB index can be obtained by averaging all the cluster similarities. The smaller the index is, the better the clustering result is. By minimizing this index, clusters are the most distinct from each other, and therefore achieves the best partition. The Xie-Beni index (XB) [13] defines the inter-cluster separation as the minimum square distance between cluster centers, and the intra-cluster compactness as the mean square distance between each data object and its cluster center. The optimal cluster number is reached when the minimum of XB is found. Kim et al. [14] proposed indices DB^{**} and

Table II
EXPERIMENT RESULTS OF THE IMPACT OF MONOTONICITY, TRUE $NC = 5$

	$RMSSTD$	RS	Γ	CH	I	D	S	DB^{**}	SD	S_Dbw	XB^{**}
2	28.496	0.627	2973	1683	3384	0.491	0.607	0.716	0.215	61.843	0.265
3	20.804	0.801	3678	2016	5759	0.549	0.707	0.683	0.124	0.153	0.374
4	14.829	0.899	4007	2968	11230	0.580	0.825	0.522	0.075	0.059	0.495
5	**3.201**	**0.994**	**4342**	**52863**	**106163**	**2.234**	**0.913**	**0.122**	**0.045**	**0.004**	**0.254**
6	3.081	0.995	4343	45641	82239	0.025	0.718	0.521	0.504	0.066	35.099
7	2.957	0.996	4344	41291	68894	0.017	0.579	0.803	0.486	0.098	35.099
8	2.834	0.996	4346	38580	58420	0.009	0.475	1.016	0.538	0.080	36.506
9	2.715	0.997	4347	36788	50259	0.010	0.391	1.168	0.553	0.113	38.008

XB^{**} in year 2005 as the improvements of DB and XB. In this paper, we will use these two improved measures.

The idea of SD index (SD) [15] is based on the concepts of the average scattering and the total separation of clusters. The first term evaluates compactness based on variances of cluster objects, and the second term evaluates separation difference based on distances between cluster centers. The value of this index is the summation of these two terms, and the optimal number of clusters can be obtained by minimizing the value of SD.

The S_Dbw index (S_Dbw) [16] takes density into account to measure the inter-cluster separation. The basic idea is that for each pair of cluster centers, at least one of their densities should be larger than the density of their midpoint. The intra-cluster compactness is the same as it is in SD. Similarly, the index is the summation of these two terms and the minimum value of S_Dbw indicates the optimal cluster number.

There are some other internal validation measures in literature [17] [18] [19] [20]. However, some have poor performance while some are designed for data sets with specific structures. Take Composed Density between and within clusters index ($CDbw$) and Symmetry distance-based index ($Sym\text{–}index$) for examples. It is hard for $CDbw$ to find the representatives for each cluster, which makes the result of $CDbw$ instable. Also $Sym\text{–}index$ can only handle data sets which are internally symmetrical. As a result, we focus on the above mentioned 11 internal validation measures in the rest of the paper. And throughout this paper, we will use the acronyms of these measures.

III. UNDERSTANDING OF INTERNAL CLUSTERING VALIDATION MEASURES

In this section, we present a detailed study of the 11 internal validation measures mentioned in Section II and investigate the validation properties of different internal validation measures in different aspects, which may be helpful for index selection. If not mentioned, we use K-means [21] (implemented by CLUTO) [22] as the clustering algorithm for experiment.

A. *The Impact of Monotonicity*

The monotonicity of different internal validation indices can be evaluated by the following experiment. We apply the K-means algorithm on the data set *Wellseparated* and get the clustering results for different number of clusters. As shown in Figure 1, *Wellseparated* is a synthetic data set composed of five well-separated clusters.

As the experiment results shown in Table II, the first three indices monotonically increases or decreases as the cluster number NC increases. On the other hand, the rest eight indices reach their maximum or minimum value as NC equals to the true cluster number. There are certain reasons for the monotonicity of the first three indices.

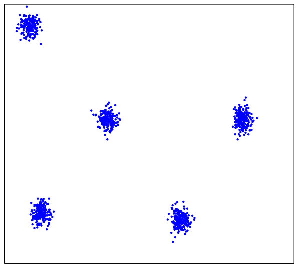


Figure 1. The Data Set *Wellseparated*

$RMSSTD = \sqrt{SSE/P(n - NC)}$, and SSE (Sum of Square Error) decreases as NC increases. In practice $NC \ll n$, thus $n - NC$ can be viewed as a constant number. Therefore, $RMSSTD$ decreases as NC increases. And we also have $RS = (TSS - SSE)/TSS$ (TSS - Total Sum of Squares), and $TSS = SSE + SSB$ (SSB - Between group Sum of Squares) which is a constant number for a certain data set. Thus, RS increases as NC increases.

From the definition of Γ, only data objects in different clusters will be counted in the equation. Therefore, if the data set is divided into two equal clusters, each cluster will have $n/2$ objects, and $n^2/4$ pairs of distances will be counted actually. If the data set is divided into three equal clusters, each cluster will have $n/3$ objects, and $n^2/3$ pairs of distances will be counted. Therefore, with the increasing of the cluster number NC, more pairs of distances are counted, which makes Γ increase.

Looking further into these three indices, we can find out that they only take either separation or compactness into account. (RS and Γ only consider separation, and $RMSSTD$ only considers compactness). As the property of monotonicity, the curves of $RMSSTD$, RS and Γ will be either upward or downward. It is claimed that the optimal cluster number is reached at the shift point of the curves, which is also known as “the elbow” [7]. However, since the judgement of the shift point is very subjective and hard to determine, we will not discuss these three indices in the further sections.

B. *The Impact of Noise*

In order to evaluate the influence of noise on internal validation indices, we have the following experiment on the data set *Wellseparated.noise*. As shown in Figure 2, *Wellseparated.noise* is a synthetic data set formulated by adding 5% noise to the data set *Wellseparated*. The cluster numbers select by indices are shown in Table III. The experiment results show that D and CH choose the wrong cluster number. From our point of view, there are certain reasons that D and CH are significantly affected by noise.

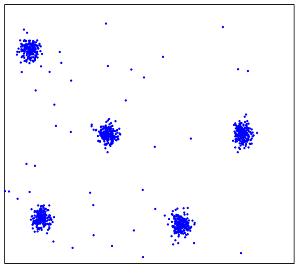


Figure 2. The Data Set *Wellseparated-noise*

Table III
EXPERIMENT RESULTS OF THE IMPACT OF NOISE, TRUE $NC = 5$

	CH	I	D	S	DB^{**}	SD	S_Dbw	XB^{**}
2	1626	3213	0.0493	0.590	0.739	0.069	20.368	0.264
3	1846	5073	0.0574	0.670	0.721	0.061	0.523	0.380
4	2554	9005	**0.0844**	0.783	0.560	0.050	0.087	0.444
5	10174	**51530**	0.0532	**0.870**	**0.183**	**0.045**	**0.025**	**0.251**
6	**14677**	48682	0.0774	0.802	0.508	0.046	0.044	0.445
7	12429	37568	0.0682	0.653	0.710	0.055	0.070	0.647
8	11593	29693	0.0692	0.626	0.863	0.109	0.052	2.404
9	11088	25191	0.0788	0.596	0.993	0.121	0.056	3.706

D uses the minimum pairwise distance between objects in different clusters ($min_{x \in C_i, y \in C_j} d(x,y)$) as the inter-cluster separation, and the maximum diameter among all clusters ($\max_k\{\max_{x,y \in C_k} d(x,y)\}$) as the intra-cluster compactness. And the optimal number of clusters can be obtained by maximizing the value of D. When noise are introduced, the inter-cluster separation can decrease sharply since it only uses the minimum pairwise distance, rather than the average pairwise distance, between objects in different clusters. Thus, the value of D may change dramatically and the corresponding optimal cluster number will be influenced by the noise.

Since $CH = (SSB/SSE) \cdot ((n - NC)/(NC - 1))$, and $((n-NC)/(NC-1))$ is constant for the same NC, we can just focus on the (SSB/SSE) part. By introducing noise, SSE increases in a more significant way comparing with SSB. Therefore, for the same NC, CH will decrease by the influence of noise, which makes the value of CH instable. Finally, the optimal cluster number will be affected by noise.

Moreover, the other indices rather than CH and D will also be influenced by noise in a less sensitive way. Comparing Table III with Table II, we can observe that the values of other indices more or less change. If we add 20% noise to the data set *Wellseparated*, the optimal cluster number suggested by I will also be incorrect. Thus, in order to minimize the adverse effect of noise, in practice it is always good to remove noise before clustering.

C. *The Impact of Density*

Data set with various density is challenging for many clustering algorithms. Therefore, we are very interested in whether it also affects the performance of the internal validation measures. An experiment is done on a synthetic data set with different density, which names *Differentdensity*. The results listed in Table IV show that only S suggests the wrong optimal cluster number. The details of *Differentdensity* is shown in Figure 3.

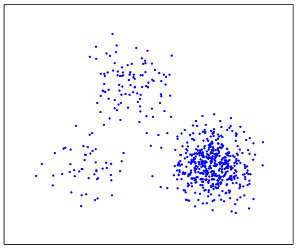


Figure 3. The Data Set *Differentdensity*

Table IV
EXPERIMENT RESULTS OF THE IMPACT OF DENSITY, TRUE $NC = 3$

	CH	I	D	S	DB^{**}	SD	S_Dbw	XB^{**}
2	1172	**120.1**	0.0493	0.587	0.658	0.705	0.603	0.408
3	**1197**	104.3	**0.0764**	**0.646**	**0.498**	**0.371**	**0.275**	**0.313**
4	1122	93.5	0.0048	0.463	1.001	0.672	0.401	3.188
5	932	78.6	0.0049	0.372	1.186	0.692	0.367	3.078
6	811	59.9	0.0049	0.312	1.457	0.952	0.312	6.192
7	734	56.1	0.0026	0.278	1.688	1.192	0.298	9.082
8	657	44.8	0.0026	0.244	1.654	1.103	0.291	8.897
9	591	45.5	0.0026	0.236	1.696	1.142	0.287	8.897

The reason why I does not give the right cluster number is not easy to tell. We can observe that I keeps decreasing as cluster number NC increases. One possible reason by our guess is the uniform effect of K-means algorithm, which tends to divide objects into relatively equal sizes [23]. I measures compactness based on the sum of distances between objects and their cluster center. When NC is small, objects with high density are likely in the same cluster, which makes the sum of distances almost remain the same. Since most of the objects are in one cluster, the total sum will not change too much. Therefore, as NC increases, I will decrease as NC is in the denominator.

D. The Impact of Subclusters

Subclusters are clusters that are closed to each other. Figure 4 shows a synthetic data set *Subcluster* which contains five clusters, and four of them are subclusters since they can form two pairs of clusters respectively.

The experiment results presented in Table V evaluate whether the internal validation measures can handle data set with subclusters. For the data set *Subcluster*, D, S, DB^{**}, SD and XB^{**} get the wrong optimal cluster numbers, while I, CH and S_Dbw suggest the correct ones. Inter-cluster separation is supposed to have a sharp decrease when cluster number changes from $NC_{optimal}$ to $NC_{optimal+1}$ [14]. However, for D, S, DB^{**}, SD and XB^{**}, sharper deceases can be observed at $NC < NC_{optimal}$. The reasons are as follows.

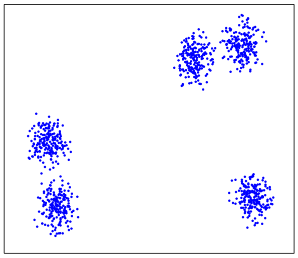


Figure 4. The Data Set *Subcluster*

Table V
Experiment Results of the Impact of Subclusters, True $NC = 5$

	CH	I	D	S	DB^{**}	SD	S_Dbw	XB^{**}
2	3474	2616	0.7410	0.736	0.445	0.156	0.207	0.378
3	7851	5008	**0.7864**	**0.803**	**0.353**	**0.096**	0.056	**0.264**
4	8670	5594	0.0818	0.737	0.540	0.164	0.039	1.420
5	**16630**	**9242**	0.0243	0.709	0.414	0.165	**0.026**	1.215
6	14310	7021	0.0243	0.587	0.723	0.522	0.063	12.538
7	12900	5745	0.0167	0.490	0.953	0.526	0.101	12.978
8	11948	4803	0.0167	0.402	1.159	0.535	0.105	14.037
9	11354	4248	0.0107	0.350	1.301	0.545	0.108	14.858

S uses the average minimum distance between clusters as the inter-cluster separation. For data set with subclusters, the inter-cluster separation will achieve its maximum value when subclusters close to each other are considered as one big cluster. Therefore, the wrong optimal cluster number will be chosen due to subclusters. XB^{**} uses the minimum pairwise distance between cluster centers as the evaluation of separation. For data set with subclusters, the measure of separation will achieve its maximum value when subclusters closed to each other are considered as a big cluster. As a result, the correct cluster number will not be found by using XB^{**}. The reasons for D, SD and DB^{**} are very similar to the reason of XB^{**}, we will not elaborate them here due to the limit of space.

E. The Impact of Skewed Distributions

It is common that clusters in a data set have unequal sizes. Figure 5 shows a synthetic data set *Skewdistribution* with skewed distributions. It consists of one large cluster and two small ones. Since K-means has the uniform effect which tends to divide objects into relatively equal sizes, it does not have a good performance when dealing with skewed distributed data sets [23]. In order to demonstrate this statement, we employ four widely used algorithms from four different categories: K-means (prototype-based), DBSCAN (density-based) [24], Agglo based on average-link (hierarchical) [2] and Chameleon (graph-based) [25]. We apply each of them on *Skewdistribution* and divide the data set into three clusters, since three is the true cluster number. As shown in Figure 6, K-means performs the worst while Chameleon is the best.

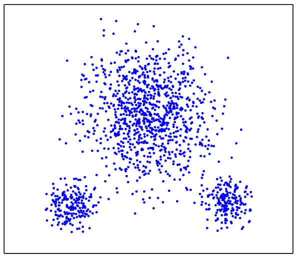


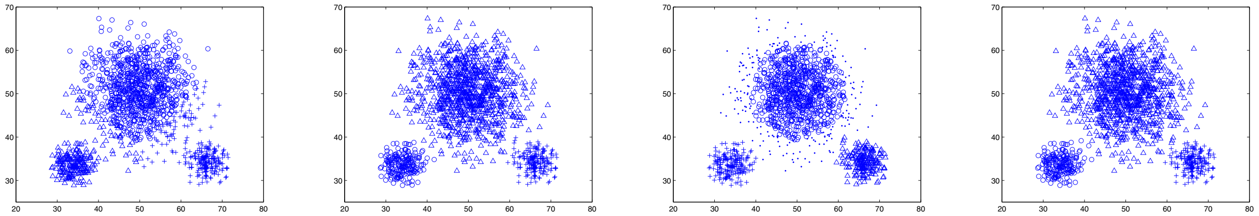
Figure 5. The Data Set *Skewdistribution*

Table VI
Experiment Results of the Impact of Skewed Distributions, True $NC = 3$

	CH	I	D	S	DB^{**}	SD	S_Dbw	XB^{**}
2	788	232.3	0.0286	0.621	0.571	0.327	0.651	0.369
3	1590	**417.9**	**0.0342**	**0.691**	**0.466**	**0.187**	**0.309**	**0.264**
4	1714	334.5	0.0055	0.538	0.844	0.294	0.379	1.102
5	**1905**	282.9	0.0069	0.486	0.807	0.274	0.445	0.865
6	1886	226.7	0.0075	0.457	0.851	0.308	0.547	1.305
7	1680	187.1	0.0071	0.371	1.181	0.478	0.378	3.249
8	1745	172.9	0.0075	0.370	1.212	0.474	0.409	3.463
9	1317	125.5	0.0061	0.301	1.875	0.681	0.398	7.716

An experiment is done on the data set *Skewdistribution* to evaluate the performance of different indices on data set with skewed distributions. We use Chameleon as the clustering algorithm. The experiment results listed in Table VI show that only CH cannot give the right optimal cluster number. Since $CH = (TSS/SSE - 1) \cdot ((n - NC)/(NC - 1))$ and TSS is a constant number of a certain data set. Thus, CH is essentially based on SSE, which shares the same basis with K-means algorithm. As mentioned above, K-means cannot handle skewed distributed data sets. Therefore, the similar conclusion can be applied to CH.

Table VII lists the validation properties of all 11 internal validation measures in all five aspects studied above, which may serve as a guidance for index selection in practice. In this table, '–' stands for property not tested, and '$\times$' denotes situation cannot be handled. From Table VII we can see, S_Dbw is the only internal validation measure which performs well in all five aspects, while the other measures have certain limitations in different scenarios, mainly in aspects of noise and subclusters.



(a) Clustering by K-means (b) Clustering by Agglo (c) Clustering by DBSCAN (d) Clustering by Chameleon

Figure 6. Clustering results on the data set *Skewdistribution* by different algorithms where NC = 3

Table VII
OVERALL PERFORMANCE OF DIFFERENT INDICES

	Mono.	Noise	Dens.	Subc.	Skew Dis.
$RMSSTD$	×	–	–	–	–
RS	×	–	–	–	–
Γ	×	–	–	–	–
CH		×			×
I			×		
D		×		×	
S				×	
DB^{**}				×	
SD				×	
S_Dbw					
XB^{**}				×	

IV. CONCLUDING REMARKS

In this paper, we investigated the validation properties of a suite of 11 existing internal clustering validation measures for crisp clustering in five different aspects: monotonicity, noise, density, subclusters and skewed distributions. Computational experiments on five synthetic data sets, which well represent the above five aspects respectively, were used to evaluate the 11 validation measures. As demonstrated by the experiment results, most of the existing measures have certain limitations in different application scenarios. S_Dbw is the only measure that performs well in all five aspects. The summation of validation properties of these 11 internal validation measures shown in Table VII may serve as a guidance for index selection in practice.

V. ACKNOWLEDGEMENTS

This research was supported in part by National Science Foundation (NSF) via grant number CNS-0831186 and the Program for New Century Excellent Talents in University of China (NCET) via grant number NCET-05-0097.

REFERENCES

[1] U. Maulik and S. Bandyopadhyay, “Performance evaluation of some clustering algorithms and validity indices,” *IEEE PAMI*, vol. 24, pp. 1650–1654, 2002.

[2] A. K. Jain and R. C. Dubes, *Algorithms for clustering data*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1988.

[3] J. Wu, H. Xiong, and J. Chen, “Adapting the right measures for k-means clustering,” in *KDD*, 2009, pp. 877–886.

[4] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. USA: Addison-Wesley Longman, Inc., 2005.

[5] Y. Zhao and G. Karypis, “Evaluation of hierarchical clustering algorithms for document datasets,” in *Procedings of CIKM*, 2002, pp. 515–524.

[6] S. Sharma, *Applied multivariate techniques*. New York, NY, USA: John Wiley & Sons, Inc., 1996.

[7] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, “On clustering validation techniques,” *Journal of Intelligent Information Systems*, vol. 17, no. 2-3, pp. 107–145, 2001.

[8] L. Hubert and P. Arabie, “Comparing partitions,” *Journal of Classification*, vol. 2, no. 1, pp. 193–218, 1985.

[9] T. Calinski and J. Harabasz, “A dendrite method for cluster analysis,” *Comm. in Statistics*, vol. 3, no. 1, pp. 1–27, 1974.

[10] J. Dunn, “Well separated clusters and optimal fuzzy partitions,” *J. Cybern.*, vol. 4, no. 1, pp. 95–104, 1974.

[11] P. Rousseeuw, “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis,” *J. Comput. Appl. Math.*, vol. 20, no. 1, pp. 53–65, 1987.

[12] D. Davies and D. Bouldin, “A cluster separation measure,” *IEEE PAMI*, vol. 1, no. 2, pp. 224–227, 1979.

[13] X. L. Xie and G. Beni, “A validity measure for fuzzy clustering,” *IEEE PAMI*, vol. 13, no. 8, pp. 841–847, 1991.

[14] M. Kim and R. S. Ramakrishna, “New indices for cluster validity assessment,” *Pattern Recogn. Lett.*, vol. 26, no. 15, pp. 2353–2363, 2005.

[15] M. Halkidi, M. Vazirgiannis, and Y. Batistakis, “Quality scheme assessment in the clustering process,” in *PKDD*, London, UK, 2000, pp. 265–276.

[16] M. Halkidi and M. Vazirgiannis, “Clustering validity assessment: Finding the optimal partitioning of a data set,” in *ICDM*, Washington, DC, USA, 2001, pp. 187–194.

[17] S. Saha and S. Bandyopadhyay, “Application of a new symmetry-based cluster validity index for satellite image segmentation,” *IEEE GRSL*, 2002.

[18] M. Halkidi and M. Vazirgiannis, “Clustering validity assessment using multi-representatives,” in *SETN*, 2002.

[19] R. Tibshirani, G. Walther, and T. Hastie, “Estimating the number of clusters in a data set via the gap statistic,” *J. Royal Statistical Society*, vol. 63, no. 2, pp. 411–423, 2001.

[20] B. S. Y. Lam and H. Yan, “A new cluster validity index for data with merged clusters and different densities,” in *IEEE ICSMC*, 2005, pp. 798–803.

[21] J. MacQueen, “Some methods for classification and analysis of multivariate observations,” in *Proceedings of BSMSP*. University of California Press, 1967, pp. 281–297.

[22] G. Karypis, *Cluto —software for clustering high-dimentional datasets*. version 2.1.2, 2006.

[23] H. Xiong, J. Wu, and J. Chen, “K-means clustering versus validation measures: a data distribution perspective,” in *KDD*, New York, NY, USA, 2006, pp. 779–784.

[24] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise,” in *KDD*, 1996, pp. 226–231.

[25] G. Karypis, E.-H. S. Han, and V. Kumar, “Chameleon: Hierarchical clustering using dynamic modeling,” *Computer*, vol. 32, no. 8, pp. 68–75, 1999.