IJCSC 0973-7391 Vol 6 • Number 2 April - Sep 2015 pp. 338-350 **Impact Factor: 2.5** **Available at** www.csjournals.com

Performance Measures for Densed and Arbitrary Shaped Clusters

K.P. Agrawal, Sanjay Garg and Pinkal Patel
Computer Science and Engineering Department, Institute of Technology, Nirma University, Ahmedabad, India.
kpa229@gmail.com, gargsv@gmail.com, pinkal08cece@gmail.com

Abstract: Validation of clusters quality is a challenging task, especially when natures of cluster aredensed and arbitrary shaped. The internal and relative cluster validation indices are the most commonlypreferable approaches to validate correct formation clustering structure and clustering algorithms.Although many indices have been proposed, but literature survey revealed that, there is no work have been performed whichare used to validate clustering structure specially for densed, sparsed and arbitrary shaped clusters.This paper proposes suitability of validation indices from amongmost commonly used validation indices and thier mathematical background. Results obtained also act as a guidelines for selecting appropriatevalidation indices and validate clustering structure specially for densed, sparsed and arbitrary shaped clusters.
Keywords: Cluster Validity Index, Density Based Clustering, DBSCAN, Partitoning Based Clustering,k-means.

1. Introduction

One of the most challenging tasks in data mining is obtaining good quality the clusters.Clustering is the unsupervised machine learning technique for identifying/categorizing patterns/objects having similar nature/characteristic in given distribution.The aim of clustering process is to identify natural structure in a dataset [19]. It can be categorized as Partitional, Hierarchical, Density Based, Grid Based and Model based clustering algorithms[1][2][6]. It is widely used in variety of domains such as statistics, image processing, biology, psychology, pattern recognition, security, remote sensing etc.

In real life, most of clustering techniques are very sensitive to their input parameters and results obtained from them are also different and there are no predefined structures of clusters and henceit is difficult to identify:

i) Correct clustering techniques/algorithms for a dataset, ii) Correct clustering structure for different parameters of same clustering algorithm and iii) The number of correct clusters in a dataset.

Cluster validation indices is the solution for these problems, which havefollowing categories[4]:

1. **External Index:** Measure similarity of clusters against known class labels.(e.g. Entropy, F-measures and Randstatistic etc.)
2. **Internal Index:** Measure the goodness of clusters *without* any external information. e.g. using Sum of Squared Error (SSE) method, others (inter,intra cluster distances) etc.
3. **Relative Index:** Compare two different clustering structureusing either external or internal indices measures.

Cluster Validation Indices have major issues in validation of dense and any shape clusters: Cluster Validation Indices cannot measure the arbitrary shaped clusters because they usually choose a representative point from each cluster and they calculate distance of the representative points and calculate other parameter based on these points (for example: variance) [23].

Present paper focuses on validation of clustering structure which are densed and arbitrary shapedin nature using Internal and Relative Indices.

2. Related work

Lot of work have already been done to identify appropriate number of cluster or best partition among datasets using internal validation indices but all of them used artificial and synthetic datasets using non-hierarchical (i.e. k-means) and hierarchical clustering algorithms (i.e. agglomerative and divisive).

A. Weingessel, E. Dimitriadou and Sara Dolnicar[35] have used cluster validation indices for determining the number of clusters in high dimensional binary datasets. Authors have done experimentation on 162 binary datasets with two different clustering algorithms i.e. k-means and hard competitive learning which is based on voting criteria, in which it was required to find out ‘number of clusters’ in particular dataset, they considered maximum choice obtained from all indices. To overcome the instabilities imposed by clustering algorithms in

the evaluation of the index's performance, algorithms were applied 100 times for each scenario. Ratkowsky-Lance index is performing better.

Milligan and Cooper [36] were showed the evaluation study of thirty (30) validity indices, also called them "Stopping criteria" because these indices helped to hierarchical clustering algorithm to decide where to stop in agglomerative process. They had performed experimentation on 108 synthetic small data sets (about 50 points each) with well-separated clusters. Based on experimentation results best six indices i.e. Caliski and Harabasz, Je(2)/Je(1), C-index, Gamma and Beale were selected. However, these indices are also data dependent. Thus, the behaviour of indices may change if different data structures are used. In the study [37], cluster validation indices are used for a quantitative evaluation of clustering results. The results of this study suggested that DB index is more reliable whenever the variance on the data set is equal to 0.16, that is, whenever the real clusters present on the data are more compact, else Huberts index is more reliable.

In 2005, authors [38] published a paper comparing sixteen (16) cluster validation indices using non-hierarchical clustering algorithm i.e. k-means. In Milligan and Cooper's study [36] it was shown that Calinski and Harabasz index performed best among the sixteen tested indices, when used with hierarchical clustering. Moreover, the simulation [38] revealed that when the k-means algorithm is used, indices utilizing the worst case improvement concept (e.g., the Ray and Turi, Davis and Bouldin, and G(+) indices) performed well compared to the others. This study also suggest that the performance/reliability of indices are varying depending on clustering method, data structure (i.e. arbitrary, spherical shaped etc.) and clustering objective (i.e. natural, homogeneous, etc.) and found best six indices (i.e. the Calinsky and Harabasz, Ray and Turi, Davis and Bouldin, and G(+)) that supports to k-means algorithm.

Bernard Desgraupes [34] discussed and implemented 27 internal cluster validation indices and also developed R-package "clusterCrit". This package also includes some external cluster validation indices.

The paper published by Qinpei Zhao, MantaoXu, and PasiFränti [11] in 2009, proposed a new sum of squares based validation index (WB) for homogeneous data based on independent variables and also done effective comparison to some other commonly used indexes i.e. Ball & Hall, Calinski&Harabasz (CH), Hartigan and Xu, Dunn, DeviesBouldin (DB), Xie-Beni (XB), Bayes Information Criterion (BIC), Silhouette Coefficient (SC) using k-means algorithm and gave good prospect compared to other indices.

More recently, Olatzet. al. [39] compared 30 cluster validation indices using three different algorithms: k-means, ward and average-linkage. They used 10 synthetic datasets and 20 real datasets. The results of this study showed that the experimental factors, noise and cluster overlap had the greatest impact on cluster validation indices. These studies also suggested that all indices used to obtained robust results. Moreover, they had also shown the statistical significance analysis of results obtained from experimentation.

To evaluate clustering results or to validate clusters, a tool CVAP (Cluster Validity Analysis Platform) has been developed by Kaijun Wang, Baijie Wang, and LiuqingPeng[8].This tool provides three functionalities :i) estimation of the number of clusters in dataset ii) evaluation of clustering results and iii) performance comparison among different clustering algorithms. Moreover, it also provides many validity indices, several clustering algorithms, datasets and procedures.

Cluster validation techniques are the methods which not only determines the "correct" number of clusters in the data set but also compare the results of two different sets of different clustering algorithms to determine the better one.

In study of Maria Halkidi, YannisBatistakis and MichalisVazirgiannis[3], they were presented the basic idea to evaluate/finding the best clustering scheme using relative criteria, for example by comparing clustering structure to other clustering schemes, resulting by the same algorithm but with different parameter values or by the different algorithm but with same parameter values. They had done experimentation with RS, RMSSTD, DB, SD and S_Dbw indices using k-means and CURE clustering algorithm and used artificial datasets. Moreover, these studies were showed that application of arbitrary shaped cluster in which tradition variability criteria (variance, density and its continuity, separation) are not much effective/sufficient.

FerencKovács, CsabaLegány and Attila Babos [23] have put their sincere effort to validate clustering schema instead to validate "number of cluster" in dataset. Authors have used relative and internal cluster validation indices and also tried to identify number of indices that is validate arbitrary shaped clusters (means clustering schema) given by density based clustering algorithms. For experimentation, they have used four validation

indices i.e. Dunn, DB, SD and S_Dbw, two different clustering algorithm k-means and DBSCAN (ADensity based Spatial Clustering of Applications with Noise)[7] and three different dataset. Their experimentation said that well separated clusters (normal distribution) dataset in which all four indices performed better but Dunn and S_Dbw are best. For ring shaped clusters dataset in which DBSCAN gave correct cluster schema compared to k-means algorithm and supported indices are Dunn &S_Dbw. For arbitrary shaped clusters (close to each) in which DBSCAN gave correct cluster schema and only dunn index was support it.

3. Cluster Validation Indices

This section describes mathematical foundation and basic definitions of internal cluster validation indices [34] and relative criteria are used for comparison purpose.

3.1 Notation

Let us define dataset M as a set of N observation and d dimension i.e. N*d Where $M=\{M_1,M_2,...,M_N\}$. The dataset is assumed to be partitioned in K clusters or groups. The coordinates of M_i are the coefficients of the i-th row of dataset M. The set of observation belongs to cluster C_k or $C^{\{k\}}$ is denoted by I_k.

Let us denote by $G^{\{k\}}$ **the centre** of the observations(n_k) in the cluster $C^{\{k\}}$ and by G the centre of all the observations.

$$G^{\{k\}} = \frac{1}{n_k}\sum_{i\in I_k} M_i$$

$$G = \frac{1}{N}\sum_{i=1}^{N} M_i$$

The within-cluster dispersion, noted $WGSS^{\{k\}}$, is the trace of scatter matrix and define as

$$WGSS^{\{k\}} \text{ or } Tr(WG^{\{k\}}) = \sum_{i\in I_k} ||M_i^{\{k\}} - G^{\{k\}}||^2$$

Finally the within-cluster sum of squares WGSS is the sum of the within-cluster dispersion for all the clusters:

$$WGSS \text{ or } Tr(WG) = \sum_{k=0}^{K} WGSS^{\{k\}}$$

And within group scatter matrix is denoted by WG.

The between-cluster dispersion, noted BGSS, is define as

$$BGSS \text{ or } Tr(BG) = \sum_{k=1}^{K} n_k||G^{\{k\}} - G||^2$$

And between group scatter matrix is denoted by WG & TSS (total sum of squared) = WGSS + BGSS.

The basic idea for some indices is based on **pairs of points** in which one can try to distinguish the pairs of points belonging to the same and different cluster. The total number of pairs of distinct points in the cluster C_k are $n_k(n_k-1)/2$ which are not depends on order of points. The total number of such pairs N_W is defined as

$$N_W = \sum_{k=1}^{K} \frac{n_k(n_k-1)}{2}$$

The total number of pairs of distinct points in the dataset is defined as

$$N_T = \frac{N(N-1)}{2}$$

Let us $N_T = N_W + N_B$ where N_B is the number of pairs constituted of points which do not belong to the same cluster and it is defined as

$$N_B = \sum_{k<k'} n_k n_{k'}$$

Let us denote by I_B the set of the N_B pairs of between cluster indices and I_W the set of the N_W pairs of within cluster indices.

3.2 Definition of Indices

Index	Definition	Journal name, Date and Ref.
Ball-Hall (↑)	The basic idea of this index is mean dispersion of clusters and define as $\text{Ball\_Hall} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_k}\text{WGSS}^{\{k\}}$	Technical Report of Stanford Research Institute, 1965 [10]
BanfeldRaftery (↓)	This index is able to handle noise, non-gaussian distribution type data and some but not all features (Orientation, size and shape). It can be described as $\text{Banfeld\_Raftery} = \sum_{k=1}^{K} n_k \log\left(\frac{\text{WGSS}^{\{k\}}}{n_k}\right)$	Biometrics, 1993 [12]
C index (↓)	This index is measures of between-cluster isolation and within-cluster coherence. Moreover it is data dependent. It can be defined as $C = \frac{S_W - S_{min}}{S_{max} - S_{min}}$ Where S_W is the sum of the N_W distances between all the pairs of points inside each cluster, $S_{min}$is the sum of the N_W smallest distances between all the pairs of points in the entire dataset and $S_{max}$is the sum of the N_W largest distances between all the pairs of points in the entire dataset.	British Journal of Mathematical and Statistical Psychologie, 1976 [22]
CalinskiHarabasz (↑)	It is based on cluster separation measures and used to select an 'appropriate' number of clusters. It can be defined as $CH = \frac{BGSS/(K-1)}{WGSS/(N-K)}$	Taylor and Francis, 1974 [14]
Davies Bouldin (↓)	To measure cluster separation, this index is used. It can be described as $DB = \frac{1}{K}\sum_{k=1}^{K}\max_{k' \neq k}\left(\frac{\delta_k + \delta_{k'}}{\Delta_{kk'}}\right)$ Where $\delta_k = \frac{1}{n_k}\sum_{i \in I_k} \|M_i^{\{k\}} - G^{\{k\}}\|^2$ and $\Delta_{kk'} = \| G^{\{k'\}} - G^{k}\|$	IEEE, 1979 [15]
Det_Ratio (↓)	This index is defined like this: $\text{Det\_Ratio} = \frac{\det(BG+WG)}{\det(WG)}$	Biometrics, 1971 [32]
Dunn (↑)	It is probably one of the most used indices in comparison study. It can be defined as $\text{Dunn} = \frac{d_{min}}{d_{max}}$ Where $d_{min}$is the minimal distance between points of different cluster and d_{max} is the largest within cluster distance.	Journal of Cybernetics, 1974 [16]
Gamma (↑)	It is the index of correlation between two vectors of data with the same size. One vector is the set of distances between pairs of points and the second vector is a binary. It's range is [-1,1]. It can be described as $\text{Gamma } (\Gamma) = \frac{\sum_{c\{k\} \epsilon C}\sum_{M_i, M_j \epsilon c_k} d(M_i, M_j)}{N_W\left(\binom{N}{2} - N_W\right)}$ Where $d(M_i,M_j)$ is the number of all point pairs in M such that M_v and M_b satisfy two conditions i) M_v and M_b are in different clusters and, ii) $d(M_v, M_b) < d(M_i, M_j)$.	Journal of the American Statistical Association, 1975 [9]
G + (↓)	It can be defined as $\text{G\_plus} = \frac{2S^-}{N_T(N_T-1)}$ Where S^- represents the number of times a distance between two points do not belonging to the same cluster is greater than	Annual Review of Ecology and Systematics, 1974 [31]

	the distance between two points belonging to the same cluster.	
Generalized Dunn's Index (GDI) (↑)	The GDI indices are used total 18 (6*3) different variations of separation estimator or between cluster distance(δ) and cohesion estimator or within cluster distance(Δ). It can be defined like this: $GDI = \frac{min_{k \neq k'} \delta(C_k, C_{k'})}{max_k \Delta(C_k)}$ with $1 \leq k \leq K$ and $1 \leq k' \leq K$ There are six different variants of δ (i.e. from δ_1 to δ_6) and three different variants of Δ (i.e from Δ_1 to Δ_3) and they are described as follow: $\delta_1 = \min_{\substack{i \epsilon I_k \\ j \epsilon I_{k'}}} d(M_i, M_j)$ $\delta_2 = \max_{\substack{i \epsilon I_k \\ j \epsilon I_{k'}}} d(M_i, M_j)$ $\delta_3 = \frac{1}{n_k n_{k'}} \sum_{\substack{i \epsilon I_k \\ j \epsilon I_{k'}}} d(M_i, M_j)$ $\delta_4 = d\left(G^{\{k\}}, G^{\{k'\}}\right)$ $\delta_5 = \frac{1}{n_k + n_{k'}} \left(\sum_{i \epsilon I_k} d\left(M_i, G^{\{k\}}\right) + \sum_{j \epsilon I_{k'}} d(M_j, G^{\{k'\}})\right)$ $\delta_6 = \max\left\{ \sup_{i \epsilon I_k} \inf_{j \epsilon I_{k'}} d\left(M_i, M_j\right), \sup_{j \epsilon I_k} \inf_{i \epsilon I_{k'}} d(M_i, M_j) \right\}$ and $\Delta_1 = \max_{\substack{i,j \epsilon I_k \\ i \neq j}} d(M_i, M_j)$ $\Delta_2 = \frac{1}{n_k(n_k - 1)} \sum_{\substack{i,j \epsilon I_k \\ i \neq j}} d(M_i, M_j)$ $\Delta_3 = \frac{2}{n_k} \sum_{i \epsilon I_k} d(M_i, G^{\{k\}})$	IEEE Transactions on Systems, Man, and Cybernetics, 1998 [13]
Ksq_DetW (↑)	It is used to determine of within group dispersion of cluster and defined as $Ksq\_Detw = K^2 det(WG)$	Biometrics, 1975 [24]
Log_Det_Ratio (↓)	Using the same notations as for Det_Ratio, the Log_Det_Ratio index is described like this: $Log\_Det\_Ratio = N \log(Det\_Ratio)$	Biometrics, 1971 [32]
Log_SS_Ratio (↓)	It is defined as $Log\_SS\_Ratio = \log(\frac{BGSS}{WGSS})$	Wiely, 1975 [21]
McClain Rao (↓)	McClain and Rao index is defined as $McClain\_Rao = \frac{AvgWGSS}{AvgBGSS}$	Journal of Marketing Research, 1975 [25]
PBM (↑)	The PBM index is based on distance between the points and their centre in clusters and the distances between the centres of clusters. $PBM = \left(\frac{1}{K} \times \frac{E_T}{E_W} \times D_B\right)^2$ Where the largest distance between two cluster centres is denoted by $D_B = \max_{k<k'} d\left(G^{\{k\}}, G^{\{k'\}}\right)$, the sum of distances between the points and their centre of each cluster is denoted by $E_W = \sum_{k=1}^{K} \sum_{i \epsilon I_k} d(M_i, G^{\{k\}})$ and the sum of distances between all points and centre of the entire dataset is defined as $E_T = \sum_{i=1}^{N} d(M_i, G)$.	Pattern Recognition, 2004 [27]
Point Biserial (↑)	It is a correlation measure between continuous variable and binary variable and defined like this: $Point\_Biserial = \frac{Cov(x,y)}{sd(x) \times sd(y)}$ Where Cov() is represented the covariance between two variable and sd() is presented the standard deviation of variable.	Psychometrika, 1981 [26]

Ray Turi (↓)	It is used for aera - colour image segmentation and also used to identify compact cluster. It is described as a quotient: $\text{Ray\_Turi} = \frac{1}{N}\frac{WGSS}{\max_{k<k'} d\left(G^{\{k\}}, G^{\{k'\}}\right)^2}$ Here bottom part of equation is the minimum of the squared distances between all the cluster centres.	Conference on Advances in Pattern Recognition and Digital Techniques, 1999 [30]
Ratkowsky Lance (↑)	The Ratkowsky Lance index is written as $\text{Ratkowsky\_Lance} = \frac{\bar{c}}{\sqrt{K}}$ Where $\bar{c}$ is the average ratio of (BGSS/TSS)$^{1/2}$ for each dimension of the data.	Australian Computer Journal, 1978 [29]
Scott Symons (↓)	The Scott-Symons index is defined like this: $\text{Scott\_Symons} = \sum_{k=1}^{K} n_k \log \det\left(\frac{\text{WG}\{k\}}{n_k}\right)$	Biometrics, 1971 [32]
SD (SD_Scat&SD_Dis)(↓)	The SD index is combination of two quantities SD_Scat (i.e for average scattering for clusters) and SD_Dis (i.e for total separation between clusters.) and it can be defined as $\text{SD} = (\alpha \cdot \text{SD\_Scat}) + \text{SD\_Dis}$ Where $\text{SD\_Scat} = \frac{1}{K}\sum_{k=1}^{K}\left\|\sigma(M^{\{k\}})\right\| / \|\sigma(M)\|$ and $\text{SD\_Dis} = \frac{D_{max}}{D_{min}}\sum_{k=1}^{K}\left(\sum_{\substack{k'=1 \\ k'\neq k}}^{K'}\left\|G^{\{k\}} - G^{\{k'\}}\right\|\right)^{-1}$ Here $D_{max}$and $D_{min}$are the largest and the smallest distance between the centre of the clusters and α is a weighting factor equal to SD_Dis obtained for the partition with the greatest number of clusters.	Journal of Intelligent Information Systems, 2001 [19]
S_Dbw (↓)	The S_Dbw index is used to evaluate compactness, separation and density based cluster but it does not work perfectly for density based cluster [15]. It is written like this: $\text{S\_Dbw} = \text{SD\_Scat} + \text{Dens\_bw}$ Where SD_Scat is similar to term used in SD index and $\text{Dens\_bw} = \frac{1}{K(K-1)}\sum_{k=1}^{K}\left(\sum_{\substack{k'=1 \\ k<k'}}^{K'}\frac{density(u_{kk'})}{max\{density(G^{\{k\}}), density(G^{\{k'\}})\}}\right)$ Here, $u_{kk'}$ is the midpoint of line segment defined by the clusters' centres $G^{\{k\}}$and $G^{\{k'\}}$.	IEEE, 2001 [20]
Silhouette (↑)	It is used to determine tightness and separation of clusters and it can be described as $\text{Silhouette} = \frac{1}{n_k}\sum_{i\epsilon I_k}\frac{b(i)-a(i)}{\max(a(i),b(i))}$ Where a(i) is the within cluster mean distance and b(i) is the smallest mean distance. They are written like $a(i) = \frac{1}{(n_k-1)}\sum_{\substack{i'\epsilon I_k \\ i'\neq i}} d(M_i, M_{i'})$ and $b(i) = \min_{k'\neq k}\left\{\frac{1}{n_{k'}}\sum_{i'\epsilon I_{k'}} d(M_i, M_{i'})\right\}$	Journal of Computational and Applied Mathematics, 1987 [28]
Tau (↑)	The Tau index that is reviewed by [23] and implemented by [34]. It can be defined as: $\text{Tau} = \frac{s^+ - s^-}{\sqrt{N_B N_W\left(\frac{N_T(N_T-1)}{2}\right)}}$ Where notation s^- is same as used for G-plus index and s^+ represents the number of times a distance between two points do not belonging to the same cluster is smaller than the distance between two points lying in the same cluster.	Annual Review of Ecology and Systematics, 1974 [31] and Psychometrika, 1981 [5]
Trace_W (↑)	The Trace_W index is used to measure within cluster dispersion. It finds the best division of cluster by analysis of variance and	Biometrika, 1965 [17]

	written like this: Trace_W = WGSS = Tr(WG)	
Trace_WiB (↑)	The Trace_WiB index is described as Trace_WiB = $Tr\left(\frac{BG}{WG}\right) = \frac{BGSS}{WGSS}$	Journal of the American Statistical Association, 1967 [18]
WemmertGancarski (↑)	It is described using quotients of distances between the points and the centres of all the clusters and defined as this: WG = $\frac{1}{N}\sum_{k=1}^{K} max\{0, n_k - \sum_{i\in I_k} R(M_i)\}$ Where R(M) = $\frac{\|M-G^{\{k\}}\|}{\min_{k'\neq k}\|M-G^{\{k'\}}\|}$	R-package, 2013 [34]
XieBeni (↓)	The XieBeni index is based on concept of compactness and separation. Moreover, it is used in fuzzy clustering but it is also apply on crisp clustering [26]. It can be defined like this: Xie_Beni = $\frac{1}{N}\frac{WGSS}{\min_{k<k'}\delta\left(C_k,C_{k'}\right)^2}$ Where $\delta(C_k, C_{k'})$ is the minimal squared distances between the points in the clusters and written like this: $\delta(C_k, C_{k'}) = \min_{\substack{i\in I_k \\ j\in I_{k'}}} d(M_i, M_j)$	IEEE, 1991 [33]

Table 1. Indices and rules for cluster validation.

*In Rule column as shown in table 1, max value for particular index indicates that maximum value is the best choice when we compared same index's results for the different algorithms with same datasets or for the same algorithm with different datasets and similarly for min value.

4. Proposed Approach

Two approaches have been adopted to validate clustering structure (which are densed and arbitrary shapedin nature):

1. Theoretical Approach
2. Practical Approach

Theoretical Approach:

The main characteristic of arbitrary shaped clusters is the representative point (i.e. center) of each cluster that is never fixed at center.From among the indices discussed in table – 1, selection of only those validation indices have been done wherecalculations arenotbased on representative point and other parameters based on these points.Based on this criterion, selected internal cluster validation indices are as follows:

1. Det_Ratio index
2. Dunn index
3. Gamma index
4. G_plus index
5. GDI (gdi11,12,21,22,31,32,51,52)
6. Ksq_DetW index
7. Log_Det_Ratio index
8. Point Biserial index
9. S_Dbw index
10. Tau index

C index is not calculated based on representative point, but it is not include in above list because it is data dependent. Moreover, it was not straight forward to incorporate the maximum or minimum within group sum of square in non-hierarchical clustering [38].

Practical Approach:

To prove theoretical approach, experimentation has been performed on classical dataset (i.e. having densed, sparse and arbitrary shaped clusters) with all internal indices described in section 3.2.

Steps for performing practical approach:

1) Select classical datasets having dense, spare and arbitrary shape in nature.
2) Select two or more clustering algorithms in which one do not identify correct clustering structure while other clustering algorithms can.
3) Select indices that supports correct clustering algorithm.
4) Compare the indices obtained from practical approach with theoretical approach.
5) Select appropriate indices which identify correct clustering structure for all given classical dataset(s).

5. Experimentation

5.1 Data Specification

For performing experimentation, we have used three different classical datasets (i.e. Jain, Spiral and Compound)having different nature[source : http://cs.joensuu.fi/sipu/datasets/] are shown in fig-1 :

1. Jain Dataset with 02 cluster and its nature (dense and any shape) is shown in figure 1(a).
2. Spiral Dataset with 03 cluster and its nature (any shape) is shown in figure 1(b).
3. Compound Dataset with 06 cluster and its nature (dense, sparse, ring and any shape with outliers) is shown in figure 1(c).

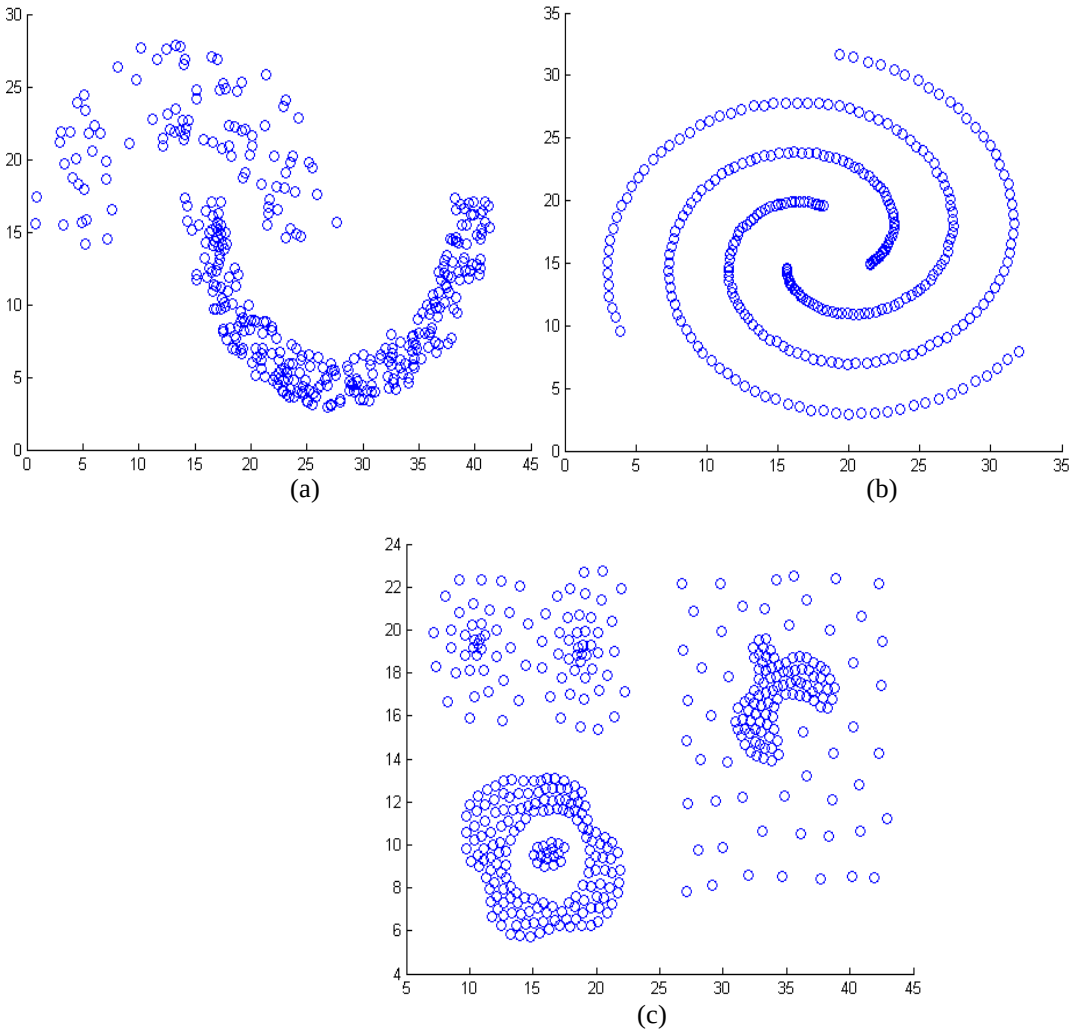


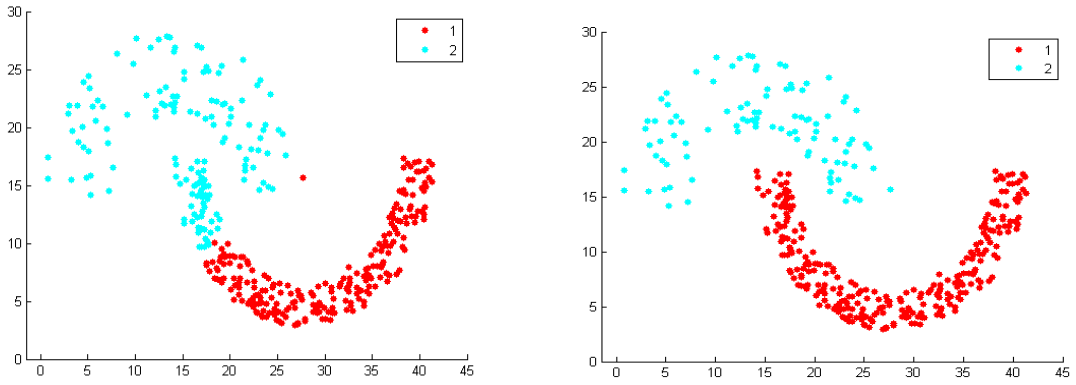
Figure 1.(a) Jain Dataset (b) Spiral Dataset and (c) Compound Dataset.

5.2 Results and Discussion

5.2.1 Practical Approach to Select appropriate validation indices

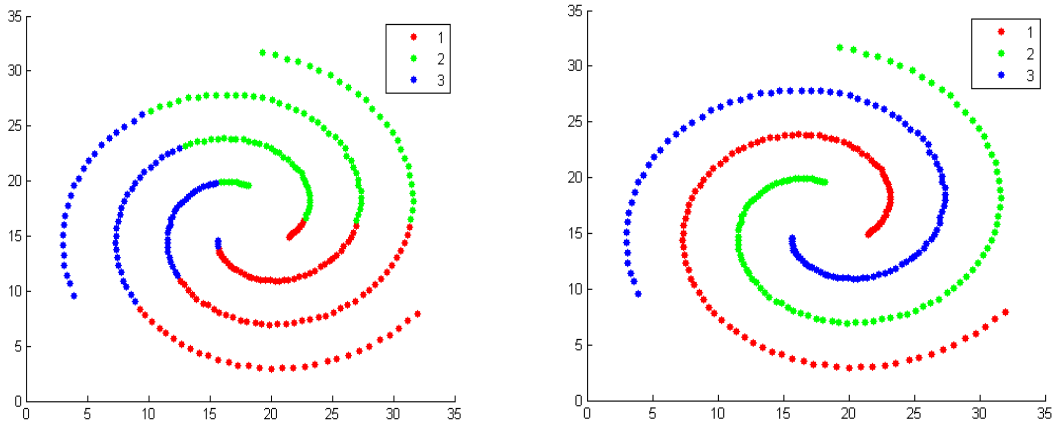
For classical dataset, following clusteringalgorithms are used:

1. k-means (Partitional Algorithm)
2. DBSCAN (Density based algorithm)



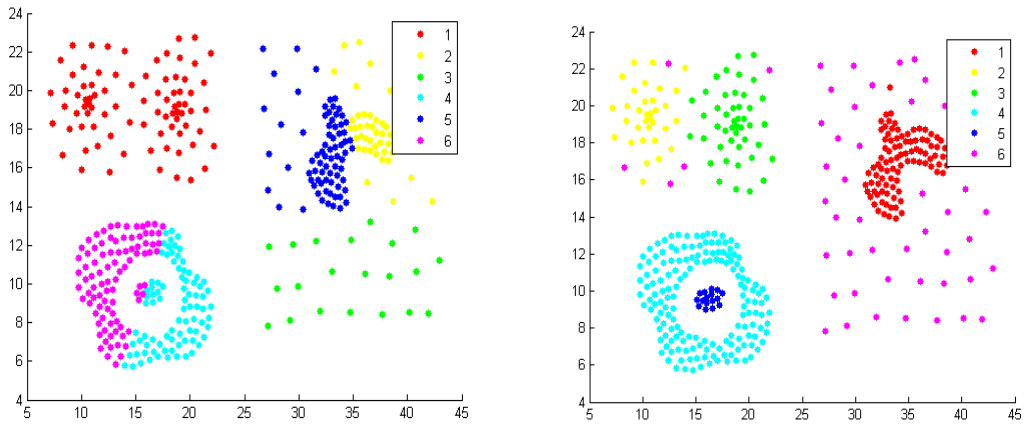
(a) (b)

Figure 2. Results of (a) k-means (b) DBSCAN on Jain dataset.



(a) (b)

Figure 3. Results of (a) k-means (b) DBSCAN on Spiral dataset.



(a) (b)

Figure 4. Results of (a) k-means (b) DBSCAN on Compound dataset.

Results of boththese algorithms for Jain, Spiral and Compound datasets have been shown in figure 2, 3 and 4 respectively. From figures 2, 3 and 4, it is clear that the results (clustering structure) of all three datasets using DBSCAN algorithm are much better than k-means algorithm.

Indices	**Rules**	**Jain**		**spiral**		**Compound**	
		DBSCAN	K-means	DBSCAN	K-means	DBSCAN	K-means
ball_hall	Max	**74.803333**	59.8308923	**96.53504591**	39.38572141	**20.1677008**	9.133920716
banfeld_raftery	Min	1631.7452	1522.939249	1425.656936	1146.05003	982.782114	941.1428588
c_index	Min	0.1807141	0.119262481	0.458272443	0.136240372	0.06438973	0.049867022
calinski_harabasz	Max	279.48312	503.4754674	5.797852185	238.3129233	361.906202	679.0463101
davies_bouldin	Min	0.8996695	0.78307242	5.948939631	0.894538706	4.56622113	0.34420882
det_ratio	Min	**3.3460281**	4.082517028	**1.076389026**	6.444023922	**21.3647099**	51.85237085
Dunn	Max	**0.0924237**	0.018699307	**0.141071797**	0.006667315	**0.04310513**	0.023997109
Gamma	Max	**-0.598760**	-0.734569132	**-0.01336378**	-0.693398654	**-0.84863883**	-0.8798154
g_plus	Min	**0.3788764**	0.433648422	**0.224517184**	0.375394933	0.34545416	0.335967732
gdi11	Max	**0.0924237**	0.018699307	**0.141071797**	0.006667315	**0.04310513**	0.023997109
gdi12	Max	**0.4459971**	0.093651276	**0.584360131**	0.035059924	**0.26333017**	0.165633773
gdi13	Max	**0.1447294**	0.032165271	**0.19145759**	0.012062476	**33**	0.059715173
gdi21	Max	1.4893557	1.607545413	1.106082483	1.35576615	0.22369928	0.503744322
gdi22	Max	7.1869919	8.051029809	4.581713126	7.129265458	1.36658381	3.47696349
gdi23	Max	2.3322325	2.765189793	1.50113553	2.452845937	0.47747084	1.253533449
gdi31	Max	0.7372479	0.783392929	0.494695097	0.692602063	0.11459354	0.24898925
gdi32	Max	3.5576427	3.923447369	2.049169979	3.642032194	0.70005441	1.718583206
gdi33	Max	1.1544815	1.347538995	0.671382467	1.253052495	0.24459208	0.619592797
gdi41	Max	0.6501577	0.713822242	0.121020417	0.610918185	0.01683587	0.211982271
gdi42	Max	3.1373824	3.575018227	0.501301522	3.212499377	0.10285071	1.463152208
gdi43	Max	1.0181040	1.227868254	0.164244575	1.105270395	0.03593502	0.527503449
gdi51	Max	**0.3053413**	0.278857359	**0.359959904**	0.273222817	0.04296875	0.099740607
gdi52	Max	**1.4734464**	1.396594391	**1.491057895**	1.43673597	0.26249705	0.688433464
gdi53	Max	0.4781443	0.479671377	0.48852469	0.494313476	0.09171387	0.24819771
ksq_detw	Max	**56398492**	462241649.1	**2032325883**	339473177.6	**511682683**	210828394.5
log_det_ratio	Min	**450.49970**	524.7042164	**22.96692673**	581.3037914	**1221.63446**	1575.411863
log_ss_ratio	Min	**-0.283260**	0.305332924	**-3.28270656**	0.433390518	**1.52701337**	2.156317629
mcclain_rao	Min	0.5546708	0.476216739	0.957905994	0.528246482	0.31594955	0.289991739
Pbm	Max	135.99496	196.5497618	1.290347306	56.3356262	123.49936	202.3043751
point_biserial	Max	**-4.351642**	-5.175361703	**-0.256728239**	-3.364320447	**-4.69923163**	-4.74318189
ray_turi	Min	0.2554356	0.183635578	9.747806492	0.234527926	59.8961278	0.732054837
ratkowsky_lance	Max	0.4932935	0.517825311	0.10974467	0.449354775	0.36192205	0.373282454
scott_symons	Min	2574.3994	2302.89743	2405.707588	1828.565898	1340.5641	1316.453629
sd_scat	Min	0.5821945	0.412999618	0.972330935	0.401221253	0.16003072	0.067195479
sd_dis	Min	0.1129814	0.111070935	0.500501531	0.120898548	3.6382463	0.606202327
s_dbw	Min	1.6638272	0.59332748	3.981600292	3.323213794	**2.35446996**	4.560024802
silhouette	Max	0.42473838	0.493174045	0.001217834	0.360534171	0.33772354	0.424481508
Tau	Max	**-0.412215**	-0.519422468	**-0.008895828**	-0.461702874	**-0.51881178**	-0.52601785

trace_w	Max	**29856.328**	22208.78484	**30109.35035**	12286.92822	**8330.50162**	4843.470606
trace_wib	Max	2.3460281	3.082517028	0.074988889	3.081063342	8.48908363	17.01666466
wemmert_gancarski	Max	0.415748	0.601653787	0	0.488620804	0.14250205	0.607735727
xie_beni	Min	12.640154	267.5999017	7.173709574	1969.059009	9.13717732	57.12481918

Table 3. Cluster validation indices using k-means and DBSCAN on classical datasets.

Results of all 27 indices are shown in table 3 for all three classical datasets with k-means and DBSCAN algorithms. As we know that DBSCAN algorithm is more suitable compared to k-means for obtaining dense and arbitrary shaped clusters which have been shown in figures 2, 3 and 4. Therefore the indices giving better results for DBSACN (i.e. which are suitable) have been marked in bold as shown in table 3, this enable us to select set of good indices. Indices which supportsdensed, sparsed, ring and arbitrary shaped clustering structure given by DBSCAN are as follows (i.e. total 12): Ball-Hall, Det_Ratio, Dunn, Gamma, G_plus, GDI (gdi11, 12, 13, 51, 52), Ksq_DetW, Log_Det_Ratio, Log_SS_Ratio, Point Biserial, Tau and Trace_W. Column 2 i.e. 'Rules' in table 3 shows the rules ("max" or "min") which suggests that results given by two chosen algorithm are compared and whichever algorithm is giving larger value is considered better than the other algorithm if rule is 'max' and otherwise if rule is 'min'.

We have select the validation indices from the both theoretical and practical approaches, which are identify more than one correct clustering structure in table 3. Selected indices (total 12) are as follows:Ball_Hall, Det_Ratio, Dunn, Gamma, G_plus, GDI (gdi11, 12,13, 51, 52), Ksq_DetW, Log_Det_Ratio, Log_SS_Ratio, Point Biserial, Tau and Trace_W.

6. Conclusions and Future work

In this paper 27 internal cluster validation indices have been discussed and summarized. Paper focuses on theoretical approach to select some validation indices that supports arbitrary shaped clusters and also provedthem practically.These validation indices have been evaluated on two different clustering algorithms (i.ek-means and DBSCAN) with different classical datasets having nature of dense, sparse and arbitrary shapedto identify their suitability.Result shows that validation indices Ball_Hall, Det_Ratio, Dunn, Gamma, G_plus, GDI (gdi11, 12, 13, 51, 52), Ksq_DetW, Log_Det_Ratio, Log_SS_Ratio, Point Biserial, Tau and Trace_W) are capable to find correct clustering structurewhich aredensed and arbitrary shaped in nature.This work can be enhanced to find applicability of these validation indices for spatial, temporal and spatio-temporal dataset in future.

Acknowledgement

This work is financially supported by Space Application Centre – Indian Space Research Organization (SAC-ISRO) under RESPOND Scheme vide NO. ISRO/RES/4/608/2013-14 dated on 22.4.13. Authorswould like to extend their sincere thanks to SAC-ISROauthorities.

References

[1] Han J., Kamber M. and Pei J. "Data Mining: Concepts and Techniques". The Morgan Kaufmann Series in Data Management Systems, Elsevier Science, 2011.

[2] S. Garg and R. C. Jain, Variations of k-mean Algorithm: A Study for High-Dimensional Large Data Sets, Information Technology Journal 5(6), pp. 1132-1135, 2006.

[3] Maria Halkidi, YannisBatistakis, and MichalisVazirgiannis. Clustering validity checking methods: part II. SIGMOD Rec. 31, 3, 19-27, 2002.

[4]Theodoridis, S. and Koutroubas, K. Pattern Recognition. Academic Press, 1999.

[5] Glenn Milligan. A Monte Carlo study of thirty internal criterion measures for cluster analysis. Psychometrika, 46:187-199, 1981.

[6] S. Garg and R. C. Jain, A Heuristic based variation of K-mean clustering algorithm for dealing with outlier, IJCSE, 4(3), pp.56-60, 2007.

[7] Ester M., Kriegel H., Sander J. and Xu. X. "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise". Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96). Portland, Oregon, 1996.

[8] Wang, K., Wang, B., &Peng, L. CVAP: Validation for Cluster Analyses. Data Science Journal, 8(May), 88–93. doi:10.2481/dsj.007-020, 2009.

[9] F. B. Baker and L. J. Hubert. Measuring the power of hierarchical cluster analysis. Journal of the American Statistical Association, 70:31-38, 1975.

[10] G. H. Ball and D. J. Hall. Isodata: A novel method of data analysis and pattern classification. Menlo Park: Stanford Research Institute. (NTIS No. AD 699616), 1965.

[11] Zhao, Q., Xu, M., &Fränti, P. Sum-of-squares based cluster validity index and significance analysis. Adaptive and Natural Computing Algorithms, 313–322, 2009.

[12] J.D. Banfield and A.E. Raftery. Model-based gaussian and non-gaussian clustering. Biometrics, 49:803-821, 1993.

[13] J. C. Bezdek and N. R. Pal. Some new indexes of cluster validity. IEEE Transactions on Systems, Man, and Cybernetics NPART B: CYBERNETICS , 28, no. 3:301-315, 1998.

[14] T. Calinski and J. Harabasz. A dendrite method for cluster analysis. Communicationsin Statistics, 3, no. 1:1-27, 1974.

[15] D. L. Davies and D. W. Bouldin. A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1, no. 2:224-227, 1979.

[16] J. Dunn. Well separated clusters and optimal fuzzy partitions. Journal of Cybernetics, 4:95-104, 1974.

[17] A. W. F. Edwards and L. Cavalli-Sforza. A method for cluster analysis. Biometrika, 56:362-375, 1965.

[18] H. P. Friedman and J. Rubin. On some invariant criteria for grouping data. Journal of the American Statistical Association, 62:1159-1178, 1967.

[19] Maria Halkidi, YannisBatistakis, and MichalisVazirgiannis. On clustering validation techniques. J. Intell. Inf. Syst., 17(2-3):107-145, 2001.

[20] M. Halkidi and M. Vazirgiannis. Clustering validity assessment: Finding the optimal partitioning of a data set. Proceedings IEEE International Conference on Data Mining, pages 187-194, 2001.

[21] J. A. Hartigan. Clustering algorithms. New York: Wiley, 1975.

[22] L. Hubert and J. Schultz. Quadratic assignment as a general data-analysis strategy. British Journal of Mathematical and Statistical Psychologie, 29:190-241, 1976.

[23] Kovács, F.,Legány, C., &Babos, A. Cluster validity measurement techniques. 6th International Symposium of Hungarian, 2005.

[24] F. H. B. Marriot. Practical problems in a method of cluster analysis. Biometrics, 27:456-460, 1975.

[25] J. O. McClain and V. R. Rao. Clustisz: A program to test for the quality of clustering of a set of objects. Journal of Marketing Research, 12:456-460, 1975.

[26] G. W. Milligan. A montecarlo study of thirty internal criterion measures for cluster analysis. Psychometrika, 46, no. 2:187-199, 1981.

[27] Bandyopadhyay S. Pakhira M. K. and Maulik U. Validity index for crisp and fuzzy clusters. Pattern Recognition, 37:487-501, 2004.

[28] Rousseeuw P.J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20:53-65, 1987.

[29] D. A. Ratkowsky and G. N. Lance. A criterion for determining the number of groups in a classification. Australian Computer Journal, 10:115-117, 1978.

[30] S. Ray and Rose H. Turi. Determination of number of clusters in k-means clustering and application in colour image segmentation. in Proceedings of the 4th International Conference on Advances in Pattern Recognition and Digital Techniques, pages 137-143, 1999.

[31] F. J. Rohlf. Methods of comparing classi_cations. Annual Review of Ecology and Systematics, 5:101-113, 1974.

[32] A. J. Scott and M. J. Symons. Clustering methods based on likelihood ratio criteria. Biometrics, 27:387-397, 1971.

[33] X.L. Xie and G. Beni. A validity measure for fuzzy clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence, 13(4):841-846, 1991.

[34] Bernard Desgraupes. "clusterCrit: Clustering Indices". R package version 1.2.3. http://CRAN.R-project.org/package=clusterCrit, 2013.

[35] Weingessel, A., Dimitriadou, E. &Dolnicar, S. An examination of indexes for determining the number of clusters in binary data sets (Working Paper 29). SFB "Adaptive Information Systems and Modeling in Economics and Management Science", 1999.

[36] Milligan, Glenn and Cooper, Martha, An examination of procedures for determining the number of clusters in a data set, Psychometrika, 50, issue 2, p. 159-179, 1985.

[37] Erika Johana Salazar G., Ana Clara Velez, Carlos Mario Parra M and Oscar Ortega L. A Cluster Validity Index for Comparing Non-hierarchical Clustering Methods. EITI, p. 1-5, 2002.

[38] Yosung Shim; Jiwon Chung; In-Chan Choi, "A Comparison Study of Cluster Validity Indices Using a Nonhierarchical Clustering Algorithm," Computational Intelligence for Modelling, Control and Automation, 2005 and International Conference on Intelligent Agents, Web Technologies and Internet Commerce, International Conference on , vol.1, no., pp.199,204, 28-30 Nov. 2005.doi: 10.1109/CIMCA.2005.1631265.

[39] Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M., &Perona, I. An extensive comparative study of cluster validity indices. Pattern Recognition, 46(1), pp. 243–256, 2013. doi:10.1016/j.patcog.2012.07.021.