5 th International Conference on Control & Signal Processing (CSP-2017)
Proceeding of Engineering and Technology –PET
Vol.26 pp.129-133

Analyzing Clustering Validation Measures based on a New Paradigm

J.C. Rojas-Thomas*[1], M. Santos #[2], N. Duro*[3], V. López#[4], M. Mora+[5]

*Informatics and Automatic Control Department, UNED
C/ Juan del Rosal s/n, 28040-Madrid, Spain
[1]correorojas@gmail.com
[3]nduro@dia.uned.es

#Computer Architecture and Automatic Control Department, University Complutense of Madrid
C/ Profesor García Santesmases 9, 28030-Madrid, Spain
[2]msantos@ucm.es
[4]vlopez@fdi.ucm.es

+Department of Computer Science, Universidad Católica del Maule, Chile
[5]mora@spock.ucm.cl

***Abstract*— This paper applies a new paradigm to define a novel internal index for clustering evaluation. This new paradigm is based on the hypothesis that in real datasets, these data may form a single, continuous cloud of points. Under this approach, density appears as the key feature to recognize clusters. According to this data model, the proposed new internal index is defined based on the degree of variability that the density presents within the clusters. In this respect, the new index is able to find the ideal partition as the one in which the variability of these internal densities of each cluster is the lowest. At the same time, in order to avoid the division of the genuine clusters, the improvement of this index through different partitions when the number of clusters increases is analyzed. Then, the partition with the largest relative improvement is selected. The proposed clustering measure has been evaluated and compared with 7 well-known indices over 17 real data sets, with very satisfactory results.**

***Keywords*— Clustering Evaluation, Internal Indexes, Density, Real Data Sets.**

I. INTRODUCTION

The process of clustering consists of classifying in an unsupervised manner a set of patterns (samples or data) in sets [1]. The goal of a clustering algorithm is to perform a partition where objects within a cluster are similar and objects grouped in other clusters are dissimilar [2]. Depending on whether the object can only be classified as belonging to a single cluster, or to several of them with different membership degrees, clustering is classified as crisp or fuzzy, respectively.

One of the most important issues in the cluster analysis is the evaluation of the results to find the partition that best fits the underlying structure of the data. This is the main objective of clustering validation [3].

Internal validation requires no extra information about the data neither the repetition of the clustering process. It depends on some properties of the resulting clusters, such as their levels of compactness, the degree of separation and the level of roundness [4].

A typical example of compactness measure is the variance. A low value of the variance is an indicator of a strong proximity between the elements of a cluster. Separation, on the other hand, shows how different two clusters are, computing the distance between them. The distance between representative objects or between their means is a good indicator [5].

Although they differ in the way they measure the inter-clusters separation and clusters compactness level, all the internal indexes for crisp clustering, based on these two concepts, assume that clusters form compact data clouds with a high degree of separation. The innovations that have been made in this area have focused mainly on working with clusters that present geometries other than spherical [6], with remarkable exceptions such as in [7], where implicit hierarchical structure and/or natural clustering dependencies have been revealed by visualizing and analyzing input/output relations. But in general the basic paradigm has not been modified.

In the literature of clustering algorithms there are a few authors who have proposed other types of clusters configurations to be analyzed. Specifically, in [8], where some different clustering problems are presented, one of them is called *"density gradient problem"*. It consists of two clusters that form two almost adjacent regions, whose main difference is not the distance between them but their respective densities (Fig. 1). This concept is also addressed in [9], where authors propose a new algorithm of clustering based on graphs that automatically detects this configuration. However, regarding the point of view of clustering validation, this type of problems has not been addressed.

A common way to compare the performance of these internal indices is by measuring how good their predictions are regarding the right number of clusters on datasets where this number is known in advance. These datasets are typically of two types: synthetically generated or real datasets.

Based on the authors' experience with real data sets [10], this paper proposes a new paradigm to define a new internal validation index. This new approach has been called the *"Continuous Region Paradigm"*, which extends the definition of [8] to include *"n"* adjacent clusters forming a single data cloud (Fig. 2).

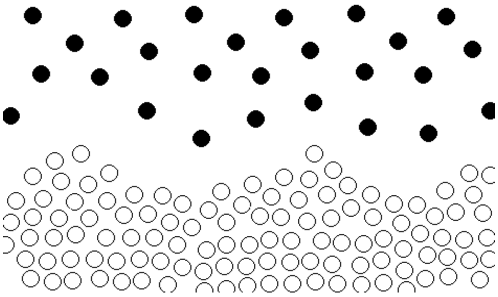

Fig. 1 “Density gradient problem” [8].

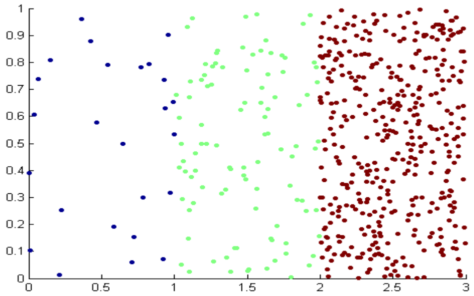

Fig. 2 Example of the continuous region paradigm on three adjacent clusters.

By comparing the performance of the proposed index with other traditional validation measures on different real data sets, this work proves that this new paradigm gives better results than the traditional one that assumes well separated compact clouds.

The organization of this paper is as follows. Section 2 briefly describes the main criteria of the internal clustering measures. Section 3 presents the definition of the proposed new index. The methodology is described in Section 4, where the experimental results are also discussed. Finally, Section 5 summarizes the main conclusions and future works.

II. Internal Validation Indices

The internal clustering validation measures are designed to reflect both compactness and separation at the same time. Naturally, considering only one of the two criteria is not enough to evaluate complex clustering. Therefore, most internal indexes are usually defined by the combination of the following two criteria:

A. Compactness

Measures how closely objects are grouped in a cluster. Compactness is normally based on distances between in-cluster points. The variance is a common way of calculating this property. A small variance indicates a high degree of compactness of the cluster.

B. Separation

Measures how different the found clusters are from each other, usually computing the distance between them. That is, how well separated they are. Separation is an inter-cluster criterion. The distance between objects is widely used because of its computational efficiency and effectiveness with clusters with hyper spherical shapes.

As a general rule, a good partition will have small intra-cluster distances and large inter-cluster distances [3].

Besides, the cluster validation indices can be classified according to the way these two criteria are combined, by the ratio between the intra-cluster and inter-cluster distances, or vice versa (Dunn, Davies Bouldin, XB, I index), or by calculating the weighted sum of these two criteria, such as the SD and S_Dbw index [11].

III. New Index Definition

Using the *Continuous Region* paradigm, a new validation index is defined in order to identify the partition that presents the biggest difference between the clusters regarding their respective densities, while maintaining the level of density variation within each cluster as low as possible.

In general terms, the strategy consists of obtaining a global estimated value of the degree of density variation in the different clusters of each partition. Then, these values are sorted in increasing order in relation to the number of clusters of their respective partitions. Finally, the partition that shows the greatest relative improvement is selected.

To define the new index, the following measures have been also defined:

A. Local Density

It estimates the local density of the neighborhood of an object x_i in the space of characteristics. It is defined as the distance to the nearest neighbor that belongs to the same cluster. It is formally expressed as (1):

$$local\_density(x_i)_{x_i \in C_k} = \min(d(x_i, x_j)), \forall x_j \in C_k, i \neq j \quad (1)$$

where *d()* corresponds to the Euclidean distance and C_k means the *k-th* cluster.

B. Density of a Cluster

It is calculated as the average value of all the local densities calculated for the data of a cluster C_k, that is, (2):

$$avg\_density(C_k) = \frac{\sum_{i=1}^{n_k} local\_density(x_i)}{n_k}, \forall x_i \in C_k \quad (2)$$

where n_k is the total number of data in cluster C_k.

C. Uniformity

It measures the degree of variation of local densities in a cluster (3):

$$uniformity(C_k)\begin{cases} n_k > 1 : \dfrac{\sum_{i=1}^{n_k} \left| local\_density(x_i) - avg\_density(C_k) \right|}{avg\_density(C_k)}, \forall x_i \in C_k \; \forall x_i \in C_k \\ n_k = 1 : 0 \end{cases} \quad (3)$$

where the lower the value the greater the uniformity.

D. *CDR Index (Contiguous Density Region)*

For a partition $P^k=\{C_1,..,C_k\}$ of a data set, with k clusters, the value of the *CDR* index is defined as (4):

$$CDR(P^k)=\frac{\sum_{i=1}^{k} n_i * uniformity(C_i)}{n_{total}} \tag{4}$$

where n_i is the total number of data in the cluster C_i, and n_{total} corresponds to the total number of data in the data set. As the goal of the index is to identify the clusters with the highest levels of uniformity regarding their densities, this index must be minimized.

E. *Searching for the Optimum*

Let be $R=\{P^1,..,P^m\}$ the set of the different partitions generated by a clustering algorithm on the *S* dataset, sorted in relation to the number of clusters (from 1 to *m* clusters). Then, the *CDR* index is calculated for each of these partitions and these new values are sorted in the same way as *R*:

$$CDR(R)=\{CDR(P^1),CDR(P^2),...,CDR(P^m)\} \tag{5}$$

An example is shown in Fig. 3. The values of the CDR index for 10 different partitions of a particular data set are represented. The horizontal axis represents the number of clusters of each partition, and the vertical axis represents the values of the index.

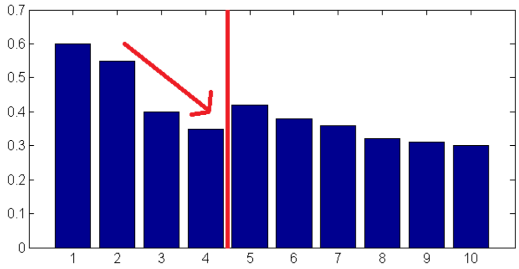

Fig. 3 Bar graph representing the searching for a local minimum in the CDR index values extracted from ten partitions of a particular data set.

In this Figure 3, the searching for a local minimum in the CDR index values is represented by the red arrow. The vertical red line represents the limit of the partitions selected for later analysis (partitions with 1 to 4 clusters).

The process of finding the minimum starts in *CDR* (P^2) (smallest solution), and continues as long as values are kept decreasing *(CDR* (P^i) *<CDR* (P^{i+1})*).* When this condition is no longer met *(CDR* (P^i)*> = CDR* (P^{i+1})*),* the process is stopped and all *i* first partitions are selected for later analysis.

The next step is to measure the degree of relative improvement between two consecutive partitions through the following factor (6):

$$Factor^k=\frac{CDR(P^k)}{CDR(P^{k-1})}, k=2,..,i \tag{6}$$

Finally, the optimal number of clusters is obtained from the partition that shows the greatest improvement (*Factor* with the lowest value) (7):

$$j: Factor^j=\min\{Factor^2,...,Factor^i\} \tag{7}$$

For this example, the *Factor* values for the previously selected partitions are shown in Table I. The lowest value obtained for the partition with 3 clusters indicates that this is the optimal number of clusters.

TABLE I
FACTOR VALUES FOR THE SELECTED PARTITIONS

K	2	3	4
Factor	0.920	0.721	0.879

IV. Methodology and Experiments

A. *Methodology*

To obtain the partitions we used the *k-means* clustering algorithm. For each data set up to 14 partitions were generated, starting with a minimum of 2 clusters up to a maximum of 15 clusters. Previously, each feature of the data was normalized to the range 0-100. For the specific case of the target partition, we do not generate a new artificial partition with the *k-means* algorithm. Instead, the target partition is directly aggregated to the set of partitions to be evaluated by the indexes.

Then, for each data set, the number of clusters selected by each internal index was recorded, considering the 14 partitions previously generated.

To compare the performance of the indices, two criteria were used:

1) *Number of hits:* number of times an index finds the correct number of clusters of the datasets.

2) *Average error:* represents the average difference, in absolute terms, between the number of clusters found by the index and the target number for a dataset. Formally it is defined as (8):

$$avg\_error=\frac{\sum_{i=1}^{n}|\text{target}-\text{prediction}|}{n} \tag{8}$$

where *target* corresponds to the target number of clusters, *prediction* to the value given by the index, and *n* is the total number of data sets analyzed.

B. *Real Data Sets*

We work with the 17 real datasets extracted from the "UCI Machine Learning Repository" [12]. These 17 real data sets are: *Iris* (3 classes, 4 features and 150 objects), *Breast Cancer Wisconsin* (Diagnostic) (2 classes, 30 features and 569 objects), *Wine* (3 classes, 13 features and 178 objects), *Vertebral Column* (3 classes, 6 features and 310 objects), *Ecoli* (8 classes, 7 features and 336 objects), *Haberman's Survival* (2 classes, 3 features and 306 objects), *Breast Tissue* (6 classes, 9 features and 106 objects), *Glass* (6 classes, 9 features and 214 objects), *Seeds* (3 classes, 7 features and 210 objects), *Spectf Heart* (2 classes, 44 features and 80 objects), *Banknote Authentication* (2 classes, 4 features, 1372 objects), *Connections Bench Sonar* (classes, 60 features, 208 objects), *Fertility* (2 classes, 9 features, 100 objects), *Parkinson* (2 classes, 22 features, 195 objects), *Statlog Vehicle* (4 classes, 18 features, 846 objects), *Yeast* (10 classes, 8 features and 1484 objects), and finally, *Steel Plates* (7 classes, 27 features and 1941 objects).

C. *Discussion of the Results*

The internal indices used in this study are: Dunn, CH, Davies-Bouldin, I, XB, Silohuette and SD.

The experiments results on the 17 real data sets are shown in Table II. Each column represents the results of a specific index, highlighted with a blue background color when the result matches the target number of clusters, and in light blue when the difference is only one unit. The last two rows represent the number of hits and the average error. As it can be seen, the new proposed CDR index shows a much higher performance than the other indices in both measures.

Analyzing the overall performance of the indices through the data sets, whenever traditional indices performed well, the same happens regarding the new internal CDR index. However, the inverse is not always true, which demonstrates the greater capacity for generalization of the new paradigm. In this sense, it is able to capture the structure of the data even when they form clouds of data with a certain degree of overlapping, showing that, in real data sets, density is a key element to differentiate classes.

V. Conclusions and Future Works

A new clustering validation measures is defined based on a novel paradigm that uses the density of the dataset as the key for generating the partitions.

The very good results obtained in the experiments with real data sets prove the effectiveness of this new proposed paradigm to evaluate partitions. In all the cases, better results than with the traditional approach have been obtained.

This inspires to continue researching on this topic, defining new internal evaluation indexes that take into account more features of the real data.

References

[1] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review", *ACM computing surveys (CSUR),* vol. 31, no 3, pp 264-323, 1999.

[2] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J.M. Pérez and I. Perona, "An extensive comparative study of cluster validity indices", *Pattern Recognition*, vol. 46, no. 1, pp. 243-256, 2013.

[3] M. Halkidi, Y. Batistakis and M. Vazirgiannis, "On clustering validation techniques", *Journal of intelligent information systems*, vol. 17, no. 2, pp. 107-145, 2001.

[4] M. Brun, C. Sima, J. Hua, J. Lowey, B. Carroll, E. Suh and E. R. Dougherty, "Model-based evaluation of clustering validation measures", *Pattern recognition*, vol. 40, no. 3, pp. 807-824, 2007.

[5] M. Kim and R. S. Ramakrishna, (2005) "New indices for cluster validity assessment", *Pattern Recognition Letters*, vol. 26, no. 15, pp. 2353-2363, 2005.

[6] J.C. Rojas Thomas, M. Santos and M. Mora, "New internal index for clustering validation based on graphs", *Expert Systems with Applications*, vol. 86, pp. 334-349, 2017

[7] A. Stolpe, A concept approach to input/output logic. Journal of Applied Logic, 13, 3, 239-258, 2015.

[8] C. T. Zahn, "Graph-theoretical methods for detecting and describing gestalt clusters", *IEEE Transactions on computers*, vol. 100, no. 1, pp. 68-86, 1971.

[9] C. Zhong, D. Miao and R. Wang, "A graph-theoretical clustering method based on two rounds of minimum spanning trees", *Pattern Recognition*, vol. 43, no. 3, pp. 752-766, 2010.

[10] J.C. Rojas Thomas, M. Mora and M. Santos, "Neural networks ensemble for automatic DNA microarray spot classification", *Neural Comput & Applic,* pp. 1-17, 2017, https://doi.org/10.1007/s00521-017-3190-6

[11] U. Maulik and S. Bandyopadhyay, "Performance evaluation of some clustering algorithms and validity indices", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 12, pp. 1650-1654, 2002.

[12] A. Asuncion and D. Newman, UCI machine learning repository. [Online]. Available: http://archive.ics.uci.edu/ml. Irvine, CA: University of California, School of Information and Computer Science, 2007.

TABLE III

Experimental results

Data Set	Target	CDR	SD	XB	Silo.	I	Dunn	CH	DB
Banknote	**2**	2	5	4	15	3	2	5	15
Breast C.	**2**	2	3	2	2	2	3	2	2
Breast T.	**6**	2	3	2	2	2	2	2	2
Columna	**3**	2	5	2	2	10	15	2	13
C.B. Sonar	**2**	3	4	3	3	3	9	3	15
Ecoli	**8**	3	4	3	3	3	3	6	3
Fertility	**2**	2	12	12	15	4	4	4	12
Glass	**6**	2	7	2	2	2	8	2	15
H. Survival	**2**	2	6	11	11	3	3	4	11
Iris	**3**	2	2	2	2	2	2	2	2
Parkinson	**2**	3	9	3	3	4	14	3	3
Seeds	**3**	4	2	2	2	2	11	2	2
S.Heart	**2**	2	15	15	13	4	12	4	13
S. Vehicle	**4**	2	2	2	2	2	10	2	2
Steel Plates	**7**	5	10	2	2	2	2	2	10
Wine	**3**	3	3	2	2	2	12	2	3
Yeast	**10**	2	3	3	2	6	11	2	11
AVG ERR		**1.77**	**3.77**	**3.94**	**4.71**	**2.53**	**5.06**	**2.35**	**5.47**
HITS		**6**	**1**	**1**	**1**	**1**	**1**	**1**	**1**