New Version of Davies-Bouldin Index for Clustering Validation Based on Cylindrical Distance

Juan Carlos Rojas Thomas
Facultad de Informática
Universidad Complutense de Madrid
Madrid, España
correorojas@gmail.com

***Abstract*—This paper presents a new version of Davies-Bouldin index for clustering validation through the use of a new distance based on density. This new distance is used as a similarity measurement between the means of the clusters, with the purpose of overcoming the limitations of the Euclidean distance. The new distance proposed allows considering the distribution of the data set and to approximate in a more accurate way the separation between clusters through the estimation of the densities along the line segments that connect the centroids.**

Keywords—*Clustering; Davies-Bouldin Index; Density; Cylindrical Distance.*

Introduction

The process of clustering consists on classifying in an unsupervised way a set of patterns (observations or data) into groups (clusters) [1]. In general, the clustering methods should search for clusters whose members are close to each other (in other words have a high degree of similarity) and well separated [2].

One of the most important issues in cluster analysis is the evaluation of clustering results to find the partition that best fits the underlying data. This is the main subject of clustering validation [2].

In general, there are three approaches to investigate clustering validation: external criteria, internal criteria and relative criteria [3].

Clustering validation approaches, which are based on relative criteria, aim at finding the best clustering scheme that a clustering algorithm can define under certain assumptions and parameter. Here the basic idea is the evaluation of a clustering structure by comparing it to other clustering schemes, resulting by the same algorithm but with different parameter values [4]. The Davies-Bouldin index falls into the latter category. Such indexes are used when the partitions generated by the applied clustering algorithm are no overlapping, meaning by this that each data belongs strictly to an only one class [4].

The Davies-Bouldin index is based on the approximately estimation of the distances between clusters and their dispersions to obtain a final value that represents the quality of the partition. One of the statistics used to estimate the distances between the clusters is the Euclidean distance between the means. The problem with this statistic is that it doesn't consider the geometry of the clusters; instead it reduces the estimation to the Euclidean distance between representative points (the means). As a result, two clusters with means very closed each other will be considered very close even if their data are not.

In order to overcome these limitations, this paper proposes a special type of distance, the cylindrical distance, which is used to calculate the distance between the means.

This distance tries to capture the data density along the straight lines that connect the means, and through this, to estimate how closed are the clusters as a whole.

This technique is validated by comparing its results with the original version of the index over real datasets.

This paper presents the following structure: first a description of the original index and its limitations and related works. Then, the cylindrical distance and the new version of the index proposed are explained. Afterword is presented the comparative performance of the original index and the new version proposed, and finally the conclusions and future works.

I. DAVIES-BOULDIN INDEX

A. Definition

This index (DB) is based on the idea that for a good partition inter cluster separation as well as intra cluster homogeneity and compactness should be high [5]. Then, to define the DB index, we need to define the dispersion measure and the cluster similarity measure [6]. In [1] the dispersion S_i of C_i cluster (1) and the separation D_{ij} between ith and jth clusters (2) are defined as:

$$S_i = \left(\frac{1}{|C_i|} \sum_{x \in C_i} D^p(x, c_i) \right)^{\frac{1}{p}}, p > 0 \quad (1)$$

Where $|C_i|$ is the number of data points in cluster C_i and c_i is the center of cluster C_i, and:

$$D_{ij} = \left(\sum_{l=1}^{d} |v_{il} - v_{jl}|^t \right)^{\frac{1}{t}}, t > 1 \quad (2)$$

Where v_i and v_j are the centroids of clusters C_i and C_j, respectively. Then, the DB index is defined as:

$$V_{DB} = \frac{1}{k}\sum_{i=1}^{k} R_i \tag{3}$$

Where k is the number of clusters and R_i is defined as:

$$R_i = \max_{i \neq j} R_{ij} \tag{4}$$

Where R_{ij} is the similarity measure between clusters C_i and C_j, and is defined as:

$$R_{ij} = \frac{S_i + S_j}{D_{ij}} \tag{5}$$

Since the goal is to achieve minimum within-cluster dispersion and maximum between-cluster separation, the number of clusters c that minimizes V_{DB} is taken as the optimal value of c [5].

B. *Limitations of the Original Proposal*

Essentially, the main limitation of the original proposal is that none of its terms considers the geometry of the spatial distribution of the clusters. This limitation is directly related to the use of the means to measure the distances between the clusters.

Specifically, the use of means as a measure of distance between clusters fails when they have differences variances, or the variances are not equal in all axes.

Using the Euclidean distance between the means to calculate the distances between the clusters can lead to situations where two clusters whose geometries are close but its means are far from each other will be considered more distant than two clusters whose geometries are more distant but its means are closer. This situation is illustrated in Fig. 1.

Another limitation of the use of the Euclidean distance between the means is its inability to distinguish when it is measuring distances between centroids that represent genuine clusters, from situations where it is in front of centroids that represent a partition that it is really dividing a genuine one. The Fig. 2 gives an example of two different configurations where the Euclidean distance between the means are equal, but actually in only one of them the centroids represents genuine well separated clusters.

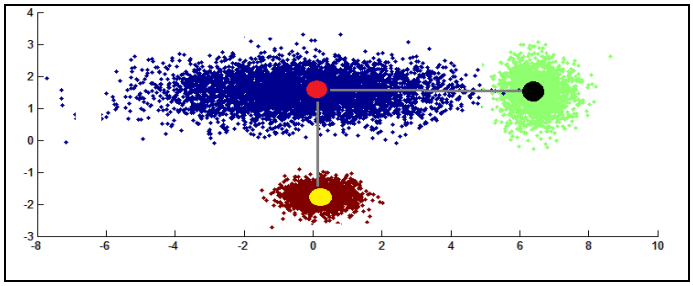

Fig. 1. The image shows two clusters with closer geometries but more distan means (blue and green clusters) and the opositte situation (blue and brown clusters).

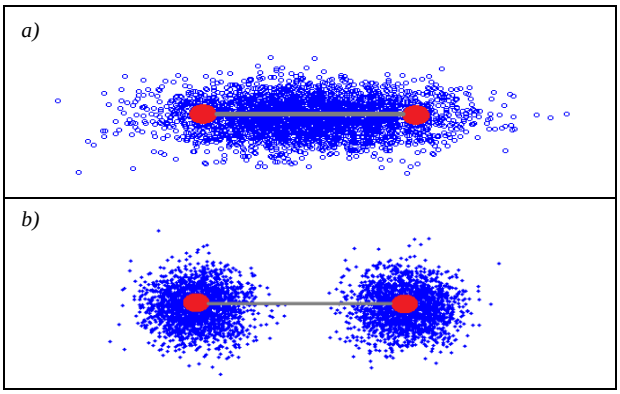

Fig. 2. The images show two configurations where the distance between the centroids are equals, but in one of them (a) the centroids belong to the same cluster, and in the other (b) the centroids belong to a well separated clusters.

A final example of the limitations of the index are ilustrated in Fig. 3. These images show two diferent configurations of clusters. In the first configuration (a) the clusters have a maximum variance vector along the horizontal axe. The second one (b) is the result of rotating the maximum variance vectors of the first configuration to a vertical position. As a result, both configurations have the same means and the same dispersions, and as a consecuence the value of the Davies-Bouldin index will be the same for both, even if in the second configuration (b) the clusters are more well separated than in the first one(a).

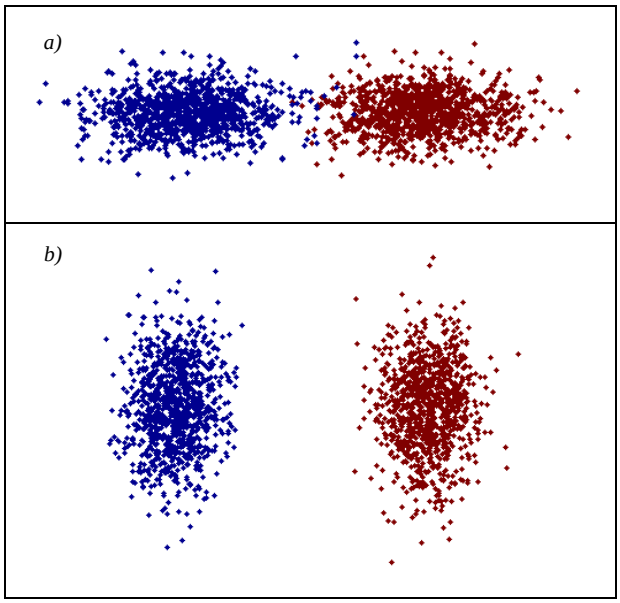

Fig. 3. The images show two configurations of clusters with the same means and dispersion, but with diferent orientations of the maximum variance vectors. Both have the same Davies-Bouldin index value.

II. RELATED WORKS

In [5] the Davies-Bouldin index is generalized through the use of graphs. Specifically, they use Minimal Spanning Tree (MST), Relative Neighborhood Graph (RNG) and Gabriel Graph (GG). However, these graphs are used only to obtain the

dispersion of each cluster, while the measurement of distance between clusters is still the distance between the means. For this reason, those approaches are not considered equivalent to this project, because they tackle another problem of the Davies-Bouldin index. Instead, they are considered complementary, and will not be used in the evaluation of the new version.

Respect to distances based on density, there is a group of clustering algorithms bases on density, where clusters are defined as dense regions separated by low-density regions [6]. Examples of these algorithms are DBSCAN [9] and BRIDGE [10]. In these algorithm the concept of density is used for determine a relation of connectivity of data points that belong to the same cluster rather than to define a new measure of distance for replacing the Euclidean one.

III. THE PROPOSAL

The proposal presented in this paper consists in a new distance, called cylindrical distance, which is used to measure the distance between the means, instead of the Euclidean one. The main idea behind this distance is to capture the data density in a limited region of the space around the straight line that connects the means, as it is illustrated in Fig. 4. This region is determined by a parameter, the radius. All data that are located in a distance, from the straight line that connects the means, equal or less that the radius, and whose projections to this segment is perpendicular, are considered to belong to this region, and used in the final calculation of the cylindrical distance. This process is illustrated in Fig 5.

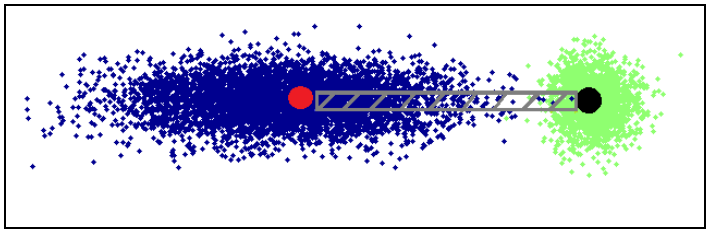

Fig. 4. The image shows an example of a region *R* generated to measure the cilyndrical distance between two centroids (red and black circles).

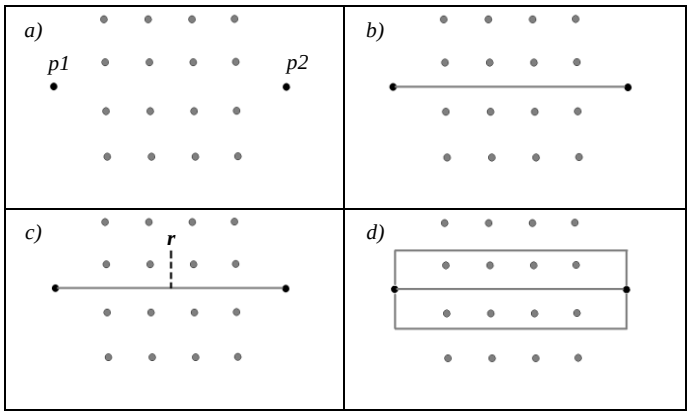

Fig. 5. The images show graphically the process of creation of the cylindrical region *R*, first (a) the two data points of which the distance is to be calculated are selected (*p1* and *p2*). Then (b) the line segment that connects both data points is built, and finally (c) using the parameter r (radius) the region is built and the data points that fall inside are used for the distance estimation (d).

A. *Definition of Cylindrical Distance*

Let *D* be the dataset of *n* dimensions.

Let *d* be any data point that belongs to the dataset *D*.

Let p_1, p_2 be the data points of which the distance is wanted to be determined.

Let $L(p_1, p_2)$ be the length of the line segment that has p_1 and p_2 as its extremes.

Let $E(d, p_1, p_2)$ be the Euclidean distance between the data point *d* and the line segment determined by the data points p_1 and p_2.

Let $R(r, p_1, p_2)$ be the region that is determined by the parameters p_1, p_2 and *r*, where p_1 and p_2 corresponds to the beginning and the ending of a line segment, and *r* the radius that determines the limits of the region *R*.

Let $P(d,p_1,p_2)$ be the angle between the line segment determined by the data points p_1 and p_2, and the shortest line that connects the data point *d* with that line segment.

Let *C* be the subset of data points that belong to the region *R*, defines as:

$$C = \{d \in D \wedge P(d, p_1, p_2) = \pi / 2 \wedge E(d, p_1, p_2) \leq r\} \quad (6)$$

Let $\rho(R)$ the *relative density* of the region $R(r, p_1, p_2)$, defined as:

$$\rho(R) = \frac{|C| + 1}{L(p_1, p_2)} \quad (7)$$

Where $|C|$ corresponds to the cardinality of the subset *C*. Then the cylindrical distance $\theta(r, p_1, p_2)$ is defined as:

$$\theta(r, p_1, p_2) = \frac{1}{\rho(R)} \quad (8)$$

Meaning that denser the region *R* is, closer will be the data points considered. As a result of the definition of the relative density of *R*, $\rho(R)$, if this region does not contain any data point, then the value of the cylindrical distance is equal to the value of the Euclidean distance between p_1 and p_2.

B. *Determination of the set of Data Points in R*

To obtain the data points that belong to the region *R* is necessary to determine first, which data points have a perpendicular projection over the line segment determined by p_1 and p_2, the data points of which the distance is to be calculated. To perform this, it is used the law of the cosines. The use of this law has the advantage that allows avoiding the complexities of working with geometries of high dimensions, and at the same time, it generalizes the algorithm to different features spaces.

To apply this law a triangle whose vertices are p_1, p_2 and the data point *d*, whose membership to the region *R* is to be determined, is built, as illustrated in Fig. 6. Then, the lengths of the sides of the triangle are calculated, and finally, the law of the cosines is applied with the purpose of obtaining the

measures of the inner angles of the triangle. If the two angles between the line segment $\overline{p_1p_2}$ and the two sides of the triangle $\overline{p_1d}$ and $\overline{p2d}$ are not greater than $\pi/2$, then the data point d has a perpendicular projection over the central line segment of the region *R*. If it is the case, the second condition to belong to this region it is the length of that projection (the line segment conformed by the data point *d* and the nearest point of the central line segment) is not greater than the parameter *r*, the radius of the region *R*.

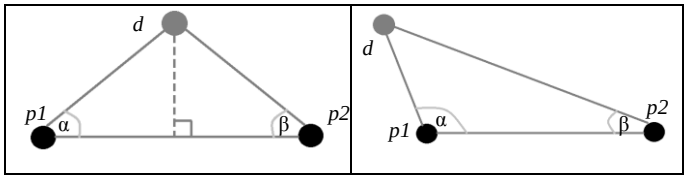

Fig. 6. The images show the two different configurations of the triangles made by the data points of which the distance is to be calculated (*p1* and *p2*) and a data point belonging to the data set (*d*). When the data point *d* has a perpendicular projection over the base of the triangle, the inner angles of the base of the triangle are always less or equal than $\pi/2$ (left image). When it is not the case, one of these angles will be always greater than $\pi/2$ (right image).

C. Cylindrical Distance Algorithm

Input: *data points p1, p2, radius r, dataset D*
Begin
n **(**Number of data points belonging to *R)* = 0
Calculate the length of $\overline{p_1p_2}$
For each Data Point *d* in *D* **Do**
Calculate the length of $\overline{p_1d}$ and $\overline{p2d}$
Calculate the angles α and β
If α<=π/2 and β<=π/2 then
Calculate $E(d, p_1, p_2)$, the Euclidean distance between *d* and $\overline{p_1p_2}$
If $E(d, p_1, p_2)$<=*r* Then
n=n+1
End If
End If
End For
θ=$L(p_1,p_2)$/(n+1)
End

D. Definition of the New Index

The new index, called the cylindrical version of Davies-Bouldin index, DB^C, is defined as:

$$DB^C = \frac{1}{k}\sum_{i=1}^{k} R_i^C \qquad (9)$$

Where *k* is the number of clusters and R_i^C is defined as:

$$R_i^C = \max_{i \neq j} R_{ij}^C \qquad (10)$$

Where R_{ij}^C is the similarity measure between clusters C_i and C_j, and is defined as:

$$R_{ij}^C = \frac{S_i + S_j}{\theta(r, v_i, v_j)} \qquad (11)$$

Where $\theta(r, v_i, v_j)$ corresponds to the cylindrical distance between the v_i, the centroid of the cluster *i*, and v_j, the centroid of the cluster *j*, and *r* to the radius of the cylindrical region. The dispersion S_i of C_i cluster is defined as usual in the traditional Davies-Bouldin index.

IV. RESULTS

A series of experiments on different sets of test were made to compare the performance of the proposed version (DB^C) with the original version of the Davies-Bouldin index. In the process of evaluation the Rand index was used, which allows to measure the level of similitude between two partitions, with values ranging from zero (minimal similitude) to one (maximal similitude) [6]. Another coefficient that was used is the Pearson correlation coefficient. This can take values from -1 to +1. A value of +1 show that the variables are perfectly linear related by an increasing relationship, a value of -1 shows that the variables are perfectly linear related by an decreasing relationship, and a value of 0 shows that the variables are not linear related by each other. It is considered a strong correlation if the correlation coefficient is greater than 0.8 and a weak correlation if the correlation coefficient is less than 0.5 [7].

The methodology used was the following: first, the ranges of values of the features of each dataset were scaled to a range of 0 to 100. Then, applying the clustering algorithm *k-means* with variable parameters were obtained 20 different partitions on each dataset used. Then, for each partition obtained, were calculated the original version of the Davies-Bouldin index (*DB*), the proposed version of the index (DB^C) using 5 different values for the radius (3, 5, 8, 10 and 15) and the Rand index, with the objective to see the similitude between the partition generated and the original classes of the data set. Finally was obtained the Pearson correlation between each version of the DB index and the Rand index considering the 20 experiments. This process was applied over 8 different datasets. They are the IRIS Dataset (3 classes, 4 features and 150 instances) [8], Ecoli Data Set (8 classes, 7 features and 336 instances) [8], Wine Data Set (3 classes, 13 features and 178 instances) [8] , Vertebral Column Data Set (3 classes, 6 features and 310 instances) [8], Climate Model Simulation Crashes Data Set (2 classes, 18 features, 540 instances) [8], Glass Identification Data Set (6 classes, 9 features, 214 instances) [8], Page Block Classification Data Set (5 classes, 10 features, 5473 instances) [8] and an Artificial Dataset of Gaussian distributions (3 classes, 2 features, 3500 instances). The Table 1 shows the values of Pearson coefficient obtained by each of the index evaluated. Because the objective of the Davies-Bouldin index and its derivatives is to be minimized, a high negative value in the Pearson coefficient indicates a good performance of the index. Those values which are highlighted indicate when the DB^C index had the best performance.

V. CONCLUSIONS AND FUTURE WORKS

In four of the datasets the proposed index had better results than the original version (Ecoli, Iris, Page and Artificial), showing a very good performance. In other two datasets where the original Davies-Bouldin index showed a good performance too (Wine and Climate) the results were equivalent, even with

the different radius used in the new version. And in the remaining two datasets where the original version of the index showed a bad performance (Column and Glass) the new version was not able to improve it. As a conclusion, the results are very promising, but, at the same time, they suggest that a more deep analysis is necessary for understanding the behavior of the index, the new version and the original one, their relations with the structural characteristics of the datasets and the influence of the clustering algorithm used for generating the partitions. An explanation of the occasions where the new index showed the same good performance than the original one could be that sometimes the structural characteristics of the datasets (and the partitions generated) make the considerations of the densities irrelevant, and in the occasions where the new index couldn't improve the bad performance of the original could be due to the structural characteristics of the datasets make impossible for the Davies-Bouldin index to obtain good results, and it would be a limitation to any new version based on it.

As a future works, it is possible to obtain better results adding modification to the cylindrical distance, with the objective of taking a more accurate measurement of the densities involved, for example, modifying the region *R* or optimizing the radius value.

TABLE I.

	DB	**DB^C**				
Radius		**3**	**5**	**8**	**10**	**15**
Wine	**-0.9360**	**-0.9360**	**-0.936**	**-0.936**	**-0.936**	**-0.936**
Ecoli	0.3317	0.1854	-0.0339	-0.1977	-0.4236	**-0.7064**
Iris	-0.3066	-0.6041	-0.8325	**-0.8676**	-0.6842	-0.4893
Column	0.9302	0.2299	0.5576	0.7824	0.9114	0.9309
Climate	**-0.8139**	**-0.8139**	**-0.8139**	**-0.8139**	**-0.8139**	**-0.8139**
Glass	0.97	0.9701	0.9648	0.7511	0.4136	0.8724
Page	0.7454	-0.7383	-0.7469	**-0.7922**	-0.7588	-0.4159
Artificial	-0.9228	**-0.9668**	-0.9495	-0.8282	-0.7365	-0.6837

REFERENCES

[1] A.K. Jain, M.N. Murty, O.J. Flynn “Data Clustering: a review”, ACM Computing Surveys, Vol. 31, No. 3, September 1999

[2] M. Halkidi, Y. Batistakis, M. Vazirgiannis, "On Clustering Validation Techniques", Intelligent Information Systems Journal, Kluwer Pulishers, 17(2-3): 107-145, 2001

[3] M. Halkidi, Y. Batistakis, M. Vazirgiannis. "Cluster Validity Methods: Part I", SIGMOD Record, June 2002

[4] M. Halkidi, Y. Batistakis, M. Vazirgiannis. "Cluster Validity Methods: Part II", SIGMOD Record, September 2002.

[5] Nikhil R. Pal, J. Biswas: Cluster validation using graph theoretic concepts. Pattern Recognition 30(6): 847-857 (1997)

[6] Guojun Gan, Chaoqun Ma, Jianhong Wu, “Data Clustering Theory, algorithms and applications” SIAM, Society for Industrial and Applied Mathematics (May 30, 2007)

[7] Sorana-Daniela BOLBOACĂ, Lorentz JÄNTSCHI, “Pearson versus Spearman, Kendall's Tau Correlation Analysis onStructure-Activity Relationships of Biologic Active Compounds”, Leonardo Journal of Sciences, Issue 9, July-December 2006, p. 179-200

[8] Frank, A. & Asuncion, A. (2010). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[9] Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining. Portland, OR, pp. 226-231.

[10] Dash, M., Liu, H. and Xu, X. (2001), "1+1>2: Merging Distance and Density Based Clustering", In Proceeding of the 7th International Conference on Database Systems for Advanced Applications (DASFAA '01), pages 32-39. Hong Kong. IEEE Computer Society Publisher.