to appear in:
Proceedings of the 14th SIAM International Conference on Data Mining (SDM), Philadelphia, PA, 2014

Density-Based Clustering Validation

Davoud Moulavi* Pablo A. Jaskowiak*† Ricardo J. G. B. Campello† Arthur Zimek‡
Jörg Sander*

Abstract

One of the most challenging aspects of clustering is validation, which is the objective and quantitative assessment of clustering results. A number of different relative validity criteria have been proposed for the validation of *globular*, clusters. Not all data, however, are composed of globular clusters. Density-based clustering algorithms seek partitions with high density areas of points (clusters, not necessarily globular) separated by low density areas, possibly containing noise objects. In these cases relative validity indices proposed for globular cluster validation may fail. In this paper we propose a relative validation index for density-based, arbitrarily shaped clusters. The index assesses clustering quality based on the relative density connection between pairs of objects. Our index is formulated on the basis of a new kernel density function, which is used to compute the density of objects and to evaluate the within- and between-cluster density connectedness of clustering results. Experiments on synthetic and real world data show the effectiveness of our approach for the evaluation and selection of clustering algorithms and their respective appropriate parameters.

1 Introduction

Clustering is one of the primary data mining tasks. Although there is no single consensus on the definition of a cluster, the clustering procedure can be characterized as the organization of data into a finite set of categories by abstracting their underlying structure, either by grouping objects in a single partition or by constructing a hierarchy of partitions to describe data according to similarities or relationships among its objects [20, 12, 18]. Over the previous decades, different clustering definitions have given rise to a number of clustering algorithms, showing a significant field development.

The variety of clustering algorithms, however, poses difficulties to users, who not only have to select the clustering algorithm best suited for a particular task, but also have to properly tune its parameters. Such choices are closely related to clustering validation, one of the most challenging topics in the clustering literature. As stated by Jain and Dubes [20], "*without a strong effort in this direction, cluster analysis will remain a black art accessible only to those true believers who have experience and great courage*". More striking than the statement itself is the fact that it still holds true after 25 years, despite all the progress that has been made.

A common approach to evaluate the quality of clustering solutions involves the use of internal validity criteria [20]. Many of such measures allow one to rank solutions accordingly to their quality and are hence called relative validity criteria. Because internal validity criteria measure clustering quality based solely on information intrinsic to the data they have great practical appeal and numerous criteria have been proposed in the literature [24, 20, 30]. The vast majority of relative validity criteria are based on the idea of computing the ratio of within-cluster scattering (compactness) to between-cluster separation. Measures that follow this definition have been designed for the evaluation of convex shaped clusters (e.g., globular clusters) and fail when applied to validate arbitrarily shaped, non-convex clusters. They are also not defined for noise objects.

Density-based clusters are found, e.g., in geographical applications, such as clusters of points belonging to rivers, roads, power lines or any connected shape in image segmentations [21]. Some attempts have been made to develop relative validity measures for arbitrarily shaped clusters [25, 16, 8, 26, 36]. As we shall see, however, these measures have serious drawbacks that limit their practical applicability. To overcome the lack of appropriate measures to the validation of density-based clusters, we propose a measure called the Density-Based Clustering Validation index (DBCV). DBCV employs the concept of Hartigan's model of density-contour trees [18] to compute the least dense region inside a cluster and the most dense region between the clusters, which are used to measure the within and between-cluster density connectedness of clusters.

Our contributions are: (i) a new core distance definition, which evaluates the density of objects w.r.t.

*Dept. of Computing Science, University of Alberta, Edmonton, AB, Canada, {moulavi, jaskowia, jsander}@ualberta.ca
†Dept. of Computer Science, University of São Paulo, São Carlos, SP, Brazil, {pablo, campello}icmc.usp.br
‡Ludwig-Maximilians-Universität München, Munich, Germany, zimek@dbs.ifi.lmu.de

other objects in the same cluster; these distances are also comparable to distances of objects inside the cluster; (ii) a new relative validity measure, based on our concept of core distance, for the validation of arbitrarily shaped clusters (along with noise, if present) and; (iii) a novel approach that makes other relative validity criteria capable of handling noise.

The remainder of the paper is organized as follows. In Section 2 we review previous attempts to tackle the validation of density-based clustering results. In Section 3 we define the problem of density-based clustering validation and introduce our relative validity index. In Section 4 we discuss aspects of the evaluation of relative validity measures and design an experimental setup to assess the quality of our index. The results of the empirical evaluation are presented in Section 5. Finally, in Section 6, we draw the main conclusions.

2 Related Work

One of the major challenges in clustering is the validation of its results, which is often described as one of the most difficult and frustrating steps of cluster analysis [20, 24]. Clustering validation can be divided into three scenarios: external, internal, and relative [20].

External clustering validity approaches such as the Adjusted Rand Index [19] compare clustering results with a pre-existing clustering (or class) structure, i.e., a ground truth solution. Although disputably useful for algorithm comparison and evaluation [13], external measures do not have practical applicability, since, according to its definition, clustering is an unsupervised task, with no ground truth solution available *a priori*.

In real world applications internal and relative validity criteria are preferred, finding wide applicability. Internal criteria measure the quality of a clustering solution using only the data themselves. Relative criteria are internal criteria able to compare two clustering structures and point out which one is better in relative terms. Although most external criteria also meet this requirement, the term relative validity criteria often refers to internal criteria that are also relative. Such a convention is adopted hereafter. There are many relative clustering validity criteria proposed in the literature [30]. Such measures are based on the general idea of computing the ratio of within-cluster scattering to between-cluster separation, with differences arising from different formulations of these two concepts.

Although relative validity measures have been successfully employed to the evaluation of globular clustering results, they are not suitable for the evaluation of arbitrarily shaped clusters, as obtained by density-based algorithms. In the density-based clustering paradigm, clusters are defined as dense areas separated by sparse regions. Therefore, clustering results can contain arbitrarily shaped clusters and noise, which such measures cannot handle properly, considering their original definition. In spite of the extensive literature on relative validation criteria, not much attention has been given to density-based clustering validation. Indeed, only a few preliminary approaches are described in the literature.

Trying to capture arbitrary shapes of clusters some authors have incorporated concepts from graph theory into clustering validation. Pal and Biswas [25] build graphs (Minimum Spanning Trees) for each clustering solution, and use information from their edges to reformulate relative measures such as Dunn [10]. Although the use of a graph can, in principle, capture arbitrary cluster structures, the measures introduced by the authors still compute compactness and separation based on Euclidean distances, favoring globular clusters. Moreover, separation is still based on cluster centroids, which is not appropriate for arbitrarily shaped clusters. Yang and Lee [33] employ a Proximity Graph to detect cluster borders and develop tests to verify if a clustering is invalid, possibly valid or good. The problem with this approach is that it does not result in a relative measure. Moreover, the tests employed by the authors require three different parameters from the user. Finally, in both approaches [25, 33] graphs are obtained directly from distances. No density concept is employed.

Occasionally, density-based concepts have been used for clustering validation. Chou et al. [8] introduce a measure that combines concepts from Dunn [10] and Davies&Bouldin [9] and is aimed to deal with clusters of different densities. The measure cannot, however, handle arbitrary shapes. Pauwels and Frederix [26] introduce a measure based on the notion of cluster homogeneity and connectivity. Although their measure provides interesting results, it has two critical parameters, e.g., the user has to set K in order to compute KNN distances and obtain cluster homogeneity. The SD and S_Dbw measures [17, 14] have similar definitions, based on concepts of scattering inside clusters and separation between clusters considering the variance and distance between centroids of the clusters. Both measures cannot, however, deal with arbitrarily shaped clusters given that they consider the center of clusters in their definitions, which is not necessarily a representative point in density-based arbitrarily shaped clusters.

The measure called $CDbw$ [15, 16] is an attempt to overcome previous drawbacks. $CDbw$ is, as far as we know, the most employed relative measure for density-based validation. The approach adopted by $CDbw$ is to consider multiple representative points per cluster, instead of one, and thereby to capture the arbitrary shape of a cluster based on the spatial distribution

of such points. *CDbw* has, however, several major drawbacks related to the multiple representatives it employs. The first of these drawbacks is how to determine the number of representative points for each cluster. Given that clusters of different sizes, densities and shapes are under evaluation, employing a fixed number of representatives for all clusters does not seem the best approach. Even if a single number of representative points is employed for all clusters (as the authors suggest), this number can still be critical to the performance of the measure and is a parameter, which is, at the very least, undesirable. Assuming that a reasonable number of representative points can be defined, the representative points themselves have to be determined. Different approaches can be employed in such a task for *CDbw*, as suggested by the authors. The adoption of different approaches to find representative points can not only be seen as another parameter, but as a significant source of instability, given that two different sets of representatives generated by different approaches, can lead to different evaluations. A later minor modification of *CDbw*, introduced by Wu and Chow [32], suffers from the same drawbacks.

3 Density-Based Clustering Validation

Hartigan's model of Density Contour Trees [18] defines density-based clusters as regions of high density separated from other such regions by regions of low density. Considering such a model we can expect a good density-based clustering solution to have clusters in which the lowest density area inside each cluster is still denser than the highest density area surrounding clusters.

Relative validity measures deemed as "traditional" take into account distances to quantify cluster variance which, combined with their separation, then amounts for clustering quality. Minimizing cluster variance and separation, however, is not the objective in density-based clustering. Therefore, a relative measure for evaluation of density-based clustering should be defined by means of densities rather than by distances.

Below we introduce the Density Based Clustering Validation (DBCV) which considers both density and shape properties of clusters. To formulate DBCV we define the notion of all-points-core-distance ($a_{pts}coredist$) which is the inverse of the density of each object with respect to all other objects inside its cluster. Using $a_{pts}coredist$ we define a symmetric reachability distance (similar to the definition by Lelis and Sander [22]) which is then employed to build a Minimum Spanning Tree (MST) inside each cluster. The MST captures both the shape and density of a cluster, since it is built on the transformed space of symmetric reachability distances. Using such MSTs (one for each cluster), DBCV finds the lowest density region in each cluster and the highest density region between pairs of clusters.

In the definitions of our concepts we use the following notations. Let $\mathbf{O} = \{\mathbf{o}_1, \cdots, \mathbf{o}_n\}$ be a dataset containing n objects in the $\mathbb{R}^d$ feature space. Let **Dist** be an $n \times n$ matrix of pairwise distances $d(\mathbf{o}_p, \mathbf{o}_q)$, where $\mathbf{o}_p, \mathbf{o}_q \in \mathbf{O}$, for a given metric distance $d(\cdot,\cdot)$. Let $KNN(\mathbf{o}, i)$ be the distance between object $\mathbf{o}$ and its i^{th} nearest neighbor. Let $C = (\{C_i\}, N)$ $1 \leq i \leq l$ be a clustering solution containing l clusters and (a possibly empty) set of noise objects N, for which n_i is the size of the i^{th} cluster and n_N is the cardinality of noise.

To estimate the density of an object within its cluster, a traditional approach is to take the inverse of the threshold distance necessary to find K objects within this threshold [18, 5]. This way, however, the density of an object is based on the distance to a single point (the k^{th} nearest neighbor). As such, this density is not as robust as density estimates that consider more objects from the neighborhood, such as Gaussian kernel density estimate. Moreover, this definition introduces a parameter (K), which is not desirable in validation.

In the following we aim to propose a new, more robust, and parameterless definition of a core distance that can be interpreted as the inverse of a density estimate and be used in the definition of a mutual reachability distance. To achieve this goal such a core distance should have the following properties: first to act as a more robust density estimate it should not depend on a single point, but rather consider all the points in a cluster in a way that closer objects have a greater contribution to the density than farther objects. This is a common property in density estimate methods such as Gaussian kernel density estimation. Second, since in the definition of a mutual reachability distance [22] the core distance of an object is compared to the distances of the object to other objects in the cluster, the core distance should be comparable to these distances. Third the core distance of an object should be approximately to the distance of a K^{th} nearest neighbor where K is not too large (representing a small neighborhood of the object).

We define the core distance of an object $\mathbf{o}$ w.r.t. all other objects inside its cluster ($a_{pts}coredist$) as follows.

Definition 1. (Core Distance of an Object) *The* all-points-core-distance *(inverse of the density) of an object* $\mathbf{o}$*, belonging to cluster* C_i *w.r.t. all other* $n_i - 1$ *objects in* C_i *is defined as:*

$$a_{pts}coredist(\mathbf{o}) = \left(\frac{\sum_{i=2}^{n_i} \left(\frac{1}{KNN(\mathbf{o},i)} \right)^d}{n_i - 1} \right)^{-\frac{1}{d}} \tag{3.1}$$

In the following we show that our definition of $a_{pts}\,coredist$ has the three aforementioned properties.

The first property holds because we calculate the inverse of the *KNN* distances to the power of dimensionality in $a_{pts}\,coredist$, resulting in higher weight of the contribution of closer objects. Note that this effect could be made stronger by using the squared Euclidean distance instead of Euclidean distance as dissimilarity.

In Proposition 3.1 we show that the second property holds for $a_{pts}\,coredist$, i.e., we prove that the $a_{pts}\,coredist$ has values between the second and the last (n^{th}) *KNN* distances of the objects.

PROPOSITION 3.1. *The all-points-core-distance of each object* $\mathbf{o}$*,* $a_{pts}\,coredist(\mathbf{o})$*, is between the second and last nearest neighbor distance of that object, i.e.,*

$$KNN(\mathbf{o}, 2) \leq a_{pts}\,coredist(\mathbf{o}) \leq KNN(\mathbf{o}, n)$$

Proof. Proof is provided in supplementary material[1].

Finally, in Propositions 3.2 and 3.3 we show that the third property holds for our definition of $a_{pts}\,coredist$ in uniform distribution.

PROPOSITION 3.2. *Let* n *objects be uniformly distributed random variables in a* d*-dimensional unit hypersphere and* $\boldsymbol{o}$ *be an object in the center of this hypersphere. The core distance of* $\boldsymbol{o}$ *is:*

$$a_{pts}\,coredist(\mathbf{o}) \approx \ln(n)^{-\frac{1}{d}} \tag{3.2}$$

Proof. Proof is provided in supplementary material[1].

PROPOSITION 3.3. *For* $a_{pts}\,coredist(\mathbf{o})$*, we have:*

$$a_{pts}\,coredist(\mathbf{o}) \approx \ln(n)^{-\frac{1}{d}} \approx KNN(\mathbf{o}, j), \tag{3.3}$$

with j *being the closest natural number to* $\frac{n}{\ln(n)}$*.*

Proof. Proof is provided in supplementary material[1].

Although the core distance of object $\mathbf{o}$, $a_{pts}coredist(\mathbf{o})$, is approximately equal to $KNN(\mathbf{o}, j)$ for an uniform distribution for some $j \approx \frac{n}{\ln(n)}$ (Propositions 3.2 and 3.3), note that, when we have a distribution other than the uniform distribution, its behavior follows our first desired property. If most of the objects are close to $\mathbf{o}$, $a_{pts}\,coredist$ tends to be a smaller value. Contrarily if most of the objects are distributed far away from $\mathbf{o}$, $a_{pts}coredist$ tends to be a greater value.

In Propositions 3.2 and 3.3, Euclidean distance is assumed as dissimilarity, however, the conclusions are similar for Squared Euclidean distance.

[1] http://webdocs.cs.ualberta.ca/~joerg/projects/sdm2014

$a_{pts}\,coredist$ is used to calculate the symmetric mutual reachability distances in Def. 2, which can be seen as the distance between objects considering their density properties. In Def. 4 we define the minimum spanning tree using mutual reachability distances to capture the shape of the clusters together with density properties. These definitions are then used to find the lowest density area (density sparseness) within—and highest density area (density separation) between—clusters in Defs. 5 and 6, which are then used to define the relative validity index DBCV in Defs. 7 and 8.

DEFINITION 2. (MUTUAL REACHABILITY DISTANCE) *The* mutual reachability distance *between two objects* $\mathbf{o}_i$ *and* $\mathbf{o}_j$ *in* $\mathbf{O}$ *is defined as* $d_{mreach}(\mathbf{o}_i, \mathbf{o}_j) = \max\{a_{pts}\,coredist(\mathbf{o}_i), a_{pts}\,coredist(\mathbf{o}_j), d(\mathbf{o}_i, \mathbf{o}_j)\}$.

Note that the comparison of $a_{pts}\,coredist$ and $d(\mathbf{o}_i, \mathbf{o}_j)$ in Def. 2 is meaningful because of the properties of $a_{pts}\,coredist$ shown in Propositions 3.1 and 3.3.

DEFINITION 3. (MUTUAL REACH. DIST. GRAPH) *The* Mutual Reachability Distance Graph *is a complete graph with objects in* $\mathbf{O}$ *as vertices and the mutual reachability distance between the respective pair of objects as the weight of each edge.*

DEFINITION 4. (MUTUAL REACH. DIST. MST) *Let* O *be a set of objects and* G *be a mutual reachability distance graph. The* minimum spanning tree *(MST) of* G *is called* MST_{MRD}*.*

We present, in brief, the overall idea behind DBCV. Considering a single cluster C_i and its objects, we start by computing the $a_{pts}\,coredist$ of the objects within C_i, from which the Mutual Reachability Distances (MRDs) for all pairs of objects in C_i are then obtained. Based on the MRDs, a Minimum Spanning Tree (MST_{MRD}) is then built. This process is repeated for all the clusters in the partition, resulting in l minimum spanning trees, one for each cluster.

Based on the MSTs obtained in the previous steps, we define a density-based clustering validation index based on the following notions of density sparseness and density separation. The density sparseness of a single cluster is defined as the maximum edge of its corresponding MST_{MRD}, which can be interpreted as the area with the lowest density inside the cluster. We define the density separation of a cluster w.r.t. another cluster as the minimum MRD between its objects and the objects from the other cluster, which can be seen as the maximum density area between the cluster and the other cluster. These two definitions are then finally combined into our validity index DBCV.

Let the set of internal edges in the MST be all edges except those with one ending vertex of degree one.

Let the set of internal objects (vertices) be all objects except those with degree one. The density sparseness and separation of clusters are given by Defs. 5 and 6.

Definition 5. (Density Sparseness of a Cluster) *The* Density Sparseness of a Cluster *(DSC) C_i is defined as the maximum edge weight of the internal edges in MST_{MRD} of the cluster C_i, where MST_{MRD} is the minimum spanning tree constructed using a_{pts}coredist considering the objects in C_i.*

Definition 6. (Density Separation) *The* Density Separation of a Pair of Clusters *(DSPC) C_i and C_j, $1 \leq i, j \leq l, i \neq j$, is defined as the minimum reachability distance between the internal nodes of the MST_{MRD}s of clusters C_i and C_j.*

Now we can compute the density-based quality of a cluster as given by Definition 7. Note that, if a cluster has better density compactness than density separation we obtain positive values of the validity index. If the density inside a cluster is lower than the density that separates it from other clusters, the index is negative.

Definition 7. (Validity Index of a Cluster) We define the validity of a cluster $C_i, 1 \leq i \leq l$, as:

$$V_C(C_i) = \frac{\min\limits_{1 \leq j \leq l, j \neq i} \Big(DSPC(C_i, C_j)\Big) - DSC(C_i)}{\max\Big(\min\limits_{1 \leq j \leq l, j \neq i} \Big(DSPC(C_i, C_j)\Big), DSC(C_i)\Big)} \tag{3.4}$$

The density-based clustering validity, DBCV, is given by Def. 8. Note that, although noise is not explicitly present in our formulation, it is implicitly considered by the weighted average that takes into account the size of the cluster ($|C_i|$) and the total number of objects under evaluation, including noise, given by $|O|$ in Eq. (3.5).

Definition 8. (Validity Index of a Clustering) *The* Validity Index of the Clustering Solution $C = \{C_i\}, 1 \leq i \leq l$ *is defined as the weighted average of the Validity Index of all clusters in C.*

$$DBCV(C) = \sum_{i=1}^{i=l} \frac{|C_i|}{|O|} V_C(C_i) \tag{3.5}$$

It is easy to verify that our index produces values between -1 and $+1$, with greater values of the measure indicating better density-based clustering solutions.

4 Experimental Setup

The evaluation of a relative validity index is usually performed as follows [30, 3]: (i) several partitions are generated with different clustering algorithms; (ii) for each clustering algorithm the ability of the new measure to identify the *correct* number of clusters, as defined by the ground truth partition of each dataset, is verified. Although commonly employed, this evaluation procedure has drawbacks [30]. In brief, it quantifies the accuracy of a given relative validity criterion according to whether or not it identifies the correct *number* of clusters for a dataset, ignoring completely relative qualities of the partitions under evaluation. Although a partition may have the correct *number* of clusters, it can present an unnatural clustering, misleading the evaluation.

Since we use datasets with a known ground truth, we choose to employ a methodology that takes full advantage of external information. This methodology was introduced by Vendramin et al. [30] and has been previously employed successfully [31]. It assesses the accuracy of relative criteria by comparing their scores against those provided by an external criterion, such as, the Adjusted Rand Index (ARI) [20]. A relative criterion is considered to be better the more similar its scores are to those provided by an external criterion. Similarity, in this case, is measured by the Pearson correlation. Although this procedure is far from perfect [13], it probably is the best procedure available. The methodology is summarized as follows:

1. Given a dataset with known ground truth, generate n_π partitions with different properties by varying the parameters of one or more clustering methods.
2. Compute the values of the relative and external validity criteria for each one of the n_π partitions.
3. Compute the correlation between the vectors with the n_π relative validity measure values and the n_π external validity measure values. This correlation quantifies the accuracy of the relative validity criterion w.r.t. the external validity measure (ARI).

An important aspect in the evaluation of the relative measures for density-based clustering is how to deal with noise objects, given that partitions generated with density-based clustering algorithms may contain noise. As far as we know, DBCV is the first relative validity measure capable of handling noise. Since other relative indices do not have this capability, noise has to be handled *prior* to their application for a fair comparison. To the best of our knowledge, there is no established procedure in the literature defining how to deal with noise objects in a partition when applying a relative validity index. We see at least five possible alternatives: (i) assign all noise to a single cluster, (ii) assign each noise point to its closest cluster, (iii) assign each noise point to a singleton cluster, (iv) remove all noise points, and (v) remove all noise points with a proportional penalty.

Following approach (i), *real clusters* end up embed-

ded in an unique cluster of noise. Approach (ii) modifies the solutions under evaluation, causing other relative indices to evaluate different clustering solutions than the ones evaluated by our measure. In approach (iii), singleton clusters become close to most of the *real clusters*, resulting in a poor overall separation, which degrades the results of all measures. Just removing the noise without any penalty in approach (iv) is not a good strategy because the coverage is not considered. For instance, a solution which has one object from each cluster and all other objects as noise results in a perfect score. However penalizing lack of coverage as in approach (v) allows the measures to deal with noise in an well behaved way. Therefore we adopt this approach in our evaluation, i.e., we evaluate measures only on points in clusters and multiply the resulting score with $(|O| - |N|)/|O|$.

Note that this is the same approach adopted for DBCV (Eq. 3.5). In our implementation we use the Squared Euclidean distance since it amplifies the effect of Property 1, which helps to better score solutions with clusters at largely different scales of separation. Further details are provided in the supplementary material.

4.1 Relative Measures We compare our measure against five well-known relative measures from the literature, namely, Silhouette Width Criterion (SWC) [27], Variance Ratio Criterion (VRC) [4], Dunn [10], and Maulik-Bandyopadhyay (MB) [23]. We also evaluate *CDbw* [16], which is, as far as we know, the most employed measure for density-based validation. All measures are available in the Cluster Validity Library [29].

4.2 Clustering Algorithms During the evaluation of our measure we consider three different density-based clustering algorithms for generating partitions: (i) the well-known DBSCAN algorithm [11] (ii) the heuristic method by Sander et al. [28], referred to here as *OPTICS-AutoCluster*, which consists of the extraction of the leaf nodes of a density-based cluster tree constructed from an OPTICS reachability plot [1] also used in [7] and, (iii) HDBSCAN* [6], which produces a hierarchy of all possible DBSCAN* partitions, each of which is evaluated by the aforementioned relative validity measures. Considering parameters for such algorithms, we simulate a scenario in which the user has no idea about which values to choose, i.e., a scenario in which a relative density-based validation index is useful.

Two different parameters are needed as input for DBSCAN, *MinPts* and ϵ. In the case of *MinPts* we choose $MinPts \in \{4, 6, \ldots, 18, 20\}$. For ϵ, we obtain the minimum and maximum values of pairwise distances for each dataset and employ 1000 different values of ϵ equally distributed within this range.

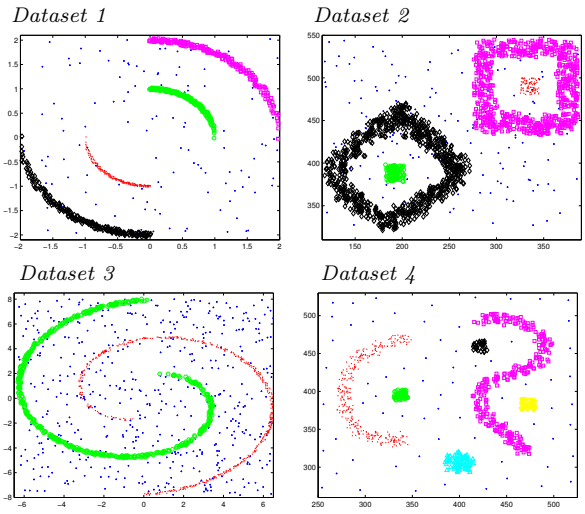


Figure 1: Synthetic 2D Datasets.

OPTICS-AutoCluster also demands *MinPts*, which was set equally to *MinPts* of DBSCAN. The speed-up control value ϵ in OPTICS was not used (ϵ = Infinity). For minimum cluster ratio we use 250 different values from 0.001 to 0.5 with steps of 0.002. Finally, for HDBSCAN* we set m_{pts} equally to *MinPts* of DBSCAN, and use, $MinClSize = m_{pts}$ as employed by its authors [6].

4.3 Datasets We employ real and synthetic datasets during our evaluation. We use real data from gene expression data sets and well-known UCI Repository [2]. We use three gene expression datasets: (i) Cell Cycle 237 (Cell237), with 237 objects, 17 features and 4 clusters; (ii) Cell Cycle 384 (Cell384), with 384 objects, 17 features and 5 clusters both were made public by [35]; and (iii) Yeast Galactose (Yeast), with 237 objects, 20 features and 4 clusters used in [34]. From UCI Repository [2], we use four datasets: (i) Iris, with 150 objects, 4 features and 3 clusters; (ii) Wine, with 178 objects, 13 features and 3 clusters; (iii) Glass, with 214 objects, 9 features and 7 clusters; and (iv) Control Chart (KDD), with 600 objects, 60 features and 6 clusters. Besides the real datasets, which are multidimensional, we also employ four 2D synthetic datasets, with different numbers of objects, clusters and noise, as depicted in Figure 1. Such datasets are useful to illustrate the behavior of our measure for arbitrarily shaped clusters.

5 Results and Discussion

5.1 Real Datasets Results for real datasets are shown in Tables 1 and 2, for which the best values for each dataset are highlighted. Table 1 shows the Adjusted Rand Index (ARI) of the *best* partition selected by each relative validity criterion. Note that DBCV

Index	Dataset						
	Cell237	Cell384	Yeast	Iris	Wine	Glass	KDD
DBCV	**0.62**	**0.39**	**0.96**	**0.60**	0.24	**0.29**	**0.56**
SWC	0.52	0.33	0.90	0.57	**0.29**	0.28	0.37
VRC	0.40	0.33	0.73	0.21	0.01	0.28	0.37
Dunn	0.35	0.16	0.38	0.13	0.01	0.28	**0.56**
CDbw	0.55	0.30	0.75	0.55	0.23	0.28	0.54
MB	0.43	0.15	0.73	0.23	0.01	0.28	**0.56**

Table 1: Best ARI found by each relative measure.

Index	Dataset						
	Cell237	Cell384	Yeast	Iris	Wine	Glass	KDD
DBCV	**0.76**	**0.79**	**0.87**	**0.97**	0.65	**0.81**	**0.84**
SWC	0.72	0.75	0.81	0.93	**0.67**	0.78	0.57
VRC	0.25	0.17	0.34	0.11	0.00	0.19	0.66
Dunn	0.64	0.29	0.65	0.25	0.10	0.62	0.51
CDbw	-0.37	-0.39	-0.06	0.83	0.59	0.09	0.01
MB	0.40	0.14	0.41	0.15	0.06	0.35	0.52

Table 2: Correlation between relative indices and ARI.

outperforms its five competitors in most of the datasets. Considering the results for the Wine dataset, in which SWC provides the best result, DBCV is a close second. For the Glass dataset, DBCV provides the best ARI value, which is the maximum obtained by all three clustering algorithms employed in the evaluation. Therefore, DBCV recognizes the best solution that is available to it. This also holds for other datasets, given that the relative measures can only find partitions as good as the ones generated by the clustering algorithms, which explains the low ARI in some cases. Table 2 shows the correlation between each relative measure and ARI. In all but one case DBCV outperforms its competitors. Again, for Wine, in which the best correlation is obtained by SWC, DBCV provides close results to SWC.

One interesting aspect that can be observed in this evaluation is that some relative measures developed for globular clusters perform relatively well. In fact, Silhouette even provides the best results for one dataset. This is easily explained by three facts. The first one is the adaptation we introduced to deal with noise for such measures, making them capable of handling partitions with noise, which can be considered an additional contribution of this paper. The second is the fact that some of the datasets employed have *globular* clusters. The third is that in some of the datasets ground truth labeling does not follow the density-based structure of the data, e.g., although ground truth consist of three globular clusters in the Iris dataset, two of these clusters are overlapping and therefore form a single cluster from a density-based perspective. In such cases, DBCV prefers 2 clusters whereas traditional measures prefer 3. The combination of these three factors makes such measures capable of recognizing good *globular* partitions in the presence of noise, as generated by density-based clustering algorithms. Note that we emphasize *globular*, since our adaption of such measures is useful only for such datasets. In case of *arbitrarily shaped* datasets, such measures are still not appropriate, as we illustrate in the following with synthetic 2D data.

Index	Dataset			
	Dataset 1	Dataset 2	Dataset 3	Dataset 4
DBCV	**0.91**	**0.90**	**0.74**	**0.99**
SWC	0.72	0.21	0.19	0.31
VRC	0.51	0.01	0.02	0.01
Dunn	0.20	0.01	0.01	0.01
CDbw	0.84	0.71	0.04	0.92
MB	0.51	0.01	0.01	0.01

Table 3: Best ARI found by each relative measure.

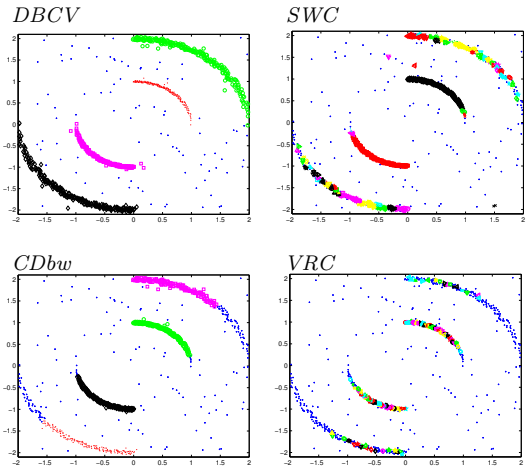
5.2 Synthetic 2D Datasets To show how the relative measures perform in the presence of *arbitrarily shaped* clusters we consider the four synthetic datasets in Figure 1. Results for this datasets are shown in Figure 2. Due to space constraints, we show plots of the best partitions selected by four relative measures, i.e., DBCV, SWC, *CDbw*, and VRC. Results for other relative measures designed for the evaluation of globular clustering solutions follow the same trend as the ones shown here. In all figures, noise points are denoted by blue dots. For all datasets, DBCV is the only measure capable of recognizing the true structure present in the data. Other relative measures like SWC and VRC often find a large number of clusters, *breaking* the true clusters into multiple small sub-clusters of small size.

As shown in Figure 2, *CDbw* is the only competitor that finds some arbitrarily shaped structure in the datasets, although it has some flaws. Considering Dataset 1, for instance, the best partition it finds has a large portion of the clusters assigned to noise (blue dots). In Dataset 2, it recognizes a clustering solution with merged clusters which is clearly not the best solution. This is also the case in Dataset 4. In Dataset 3 it is simply not capable of recognizing the best partition, which is composed of two spirals and random noise.

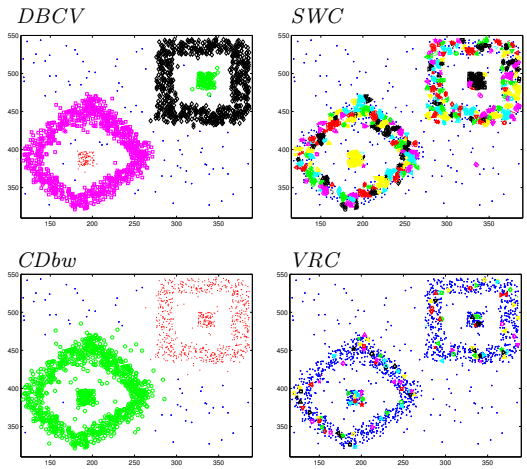
Finally, we show in Tables 3 and 4 the best ARI values found with each measure and their respective correlation with ARI, for each dataset. For all the 2D datasets, DBCV finds the best solution. It also displays the best correlation with ARI for all datasets (along with *CDbw* in dataset 4). In brief, this shows that only DBCV can properly find arbitrarily shaped clusters.

6 Conclusions

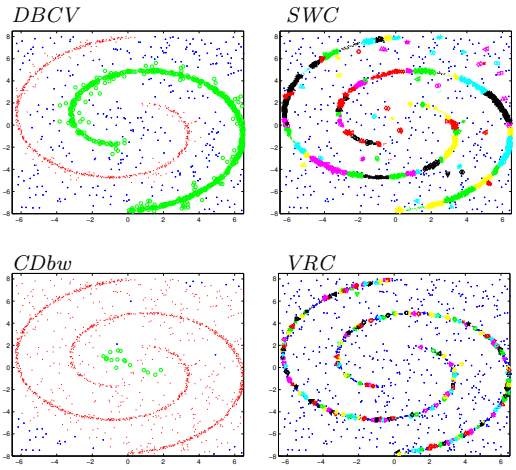
We have introduced a novel relative Density-Based Clustering Validation index, DBCV. Unlike other relative validity indices, our method not only directly takes into account density and shape properties of clusters but also properly deals with noise objects, which are intrinsic to



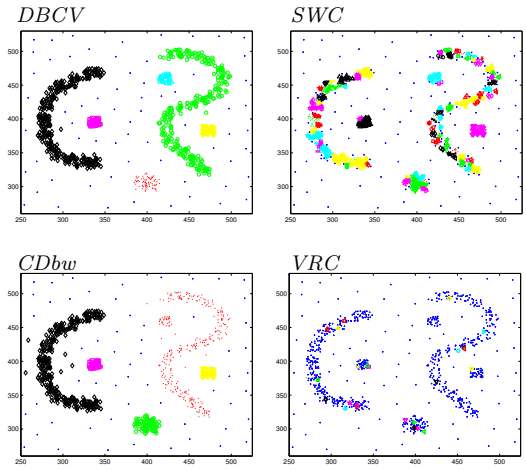
(a) Dataset 1



(b) Dataset 2



(c) Dataset 3



(d) Dataset 4

Figure 2: Best partition found on Synthetic 2D Datasets.

Index	Dataset			
	Dataset 1	Dataset 2	Dataset 3	Dataset 4
DBCV	**0.66**	**0.76**	**0.37**	**0.86**
SWC	0.39	-0.25	-0.31	-0.35
VRC	-0.15	-0.05	-0.14	-0.43
Dunn	-0.21	-0.05	-0.31	-0.32
CDbw	0.49	0.71	0.15	**0.86**
MB	-0.14	-0.16	-0.12	-0.21

Table 4: Correlation between relative indices and ARI.

the definition of the density-based clustering. We also propose an adaption to make other relative measures capable of handling noise. Both DBCV and our noise adaption approach showed promising results, confirming their efficacy and applicability to clustering validation.

Acknowledgements. This project was partially funded by NSERC (Canada). PAJ thanks Brazilian research agency FAPESP (Process #2012/15751-9). RJGBC thanks CNPq (grant #301241 2010-4) and FAPESP (grants #2010/20032−6 and #2013/18698-4).

References

[1] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander. OPTICS: Ordering points to identify the clustering structure. In *Proceedings of the ACM International Conference on Management of Data (SIGMOD), Philadelphia, PA*, pages 49–60, 1999.

[2] K. Bache and M. Lichman. UCI machine learning repository, 2013.

[3] A. B. Baya and P. M. Granitto. How many clusters: a validation index for arbitrary shaped clusters. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2013.

[4] R. B. Calinski and J. Harabasz. A dentrite method for cluster analysis. *Communications in Statistics*, 3:1–27, 1974.

[5] R. J. G. B. Campello, D. Moulavi, and J. Sander. A simpler and more accurate auto-hds framework for clustering and visualization of biological data. *IEEE/ACM TCBB*, 9(6):1850–1852, 2012.

[6] R. J. G. B. Campello, D. Moulavi, and J. Sander. Density-based clustering based on hierarchical density estimates. In *Proceedings of the 17th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Gold Coast, Australia*, 2013.

[7] R. J. G. B. Campello, D. Moulavi, A. Zimek, and J. Sander. A framework for semi-supervised and unsupervised optimal extraction of clusters from hierarchies. *Data Mining and Knowledge Discovery*, 2013.

[8] C.-H. Chou, M.-C. Su, and E. Lai. A new cluster validity measure and its application to image compression. *Pattern Analysis and Applications*, 7(2):205–220, 2004.

[9] D. Davies and D. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1:224–227, 1979.

[10] J. C. Dunn. Well separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, 4:95–104, 1974.

[11] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD96*, pages 226–231, 1996.

[12] B. S. Everitt, S. Landau, M. Leese, and D. Stahl. *Cluster Analysis*. Wiley, 2011.

[13] I. Färber, S. Günnemann, H.-P. Kriegel, P. Kröger, E. Müller, E. Schubert, T. Seidl, and A. Zimek. On using class-labels in evaluation of clusterings. In *MultiClust (with KDD) 2010, Washington, DC*, 2010.

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

[15] M. Halkidi and M. Vazirgiannis. Clustering validity assessment using multi representatives. In *Proceedings of SETN Conference*, 2002.

[16] M. Halkidi and M. Vazirgiannis. A density-based cluster validity approach using multi-representatives. *Pattern Recognition Letters*, 29:773–786, 2008.

[17] M. Halkidi, M. Vazirgiannis, and Y. Batistakis. Quality scheme assessment in the clustering process. *PKDD 2000*, page 265–276, 2000.

[18] J. A. Hartigan. *Clustering Algorithms*. John Wiley&Sons, 1975.

[19] L. Hubert and P. Arabie. Comparing partitions. *J. Classification*, 2(1):193–218, 1985.

[20] A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice Hall, 1988.

[21] H.-P. Kriegel, P. Kröger, J. Sander, and A. Zimek. Density-based clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(3):231–240, 2011.

[22] L. Lelis and J. Sander. Semi-supervised density-based clustering. In *ICDM 2009, Miami, FL*, pages 842–847, 2009.

[23] U. Maulik and S. Bandyopadhyay. Performance evaluation of some clustering algorithms and validity indices. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(12):1650–1654, 2002.

[24] G. Milligan and M. Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2):159–179, 1985.

[25] N. Pal and J. Biswas. Cluster validation using graph theoretic concepts. *Pattern Recognition*, 30(6):847 – 857, 1997.

[26] E. Pauwels and G. Frederix. Finding salient regions in images: Nonparametric clustering for image segmentation and grouping. *Computer Vision and Image Understanding*, 75(1–2):73 – 85, 1999.

[27] P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

[28] J. Sander, X. Qin, Z. Lu, N. Niu, and A. Kovarsky. Automatic extraction of clusters from hierarchical clustering representations. In *Pacific-Asia Conf. of Adv. in Knowl. Discovery and Data Mining*, pages 75–87, 2003.

[29] E. Sivogolovko and B. Novikov. Validating cluster structures in data mining tasks. In *EDBT/ICDT Workshops*, pages 245–250, 2012.

[30] L. Vendramin, R. J. G. B. Campello, and E. R. Hruschka. Relative clustering validity criteria: A comparative overview. *Statistical Analysis and Data Mining*, 3(4):209–235, 2010.

[31] L. Vendramin, P. A. Jaskowiak, and R. J. G. B. Campello. On the combination of relative clustering validity criteria. In *Proceedings of the 25th SSDBM*, page 4. ACM, 2013.

[32] S. Wu and T. W. Chow. Clustering of the self-organizing map using a clustering validity index based on inter-cluster and intra-cluster density. *Pattern Recognition*, 37(2):175 – 188, 2004.

[33] J. Yang and I. Lee. Cluster validity through graph-based boundary analysis. In *IKE*, pages 204–210, 2004.

[34] K. Yeung, M. Medvedovic, and R. Bumgarner. Clustering gene-expression data with repeated measurements. *Genome Biol.*, 4(5), 2003.

[35] K. Y. Yeung, C. Fraley, A. Murua, A. E. Raftery, and W. L. Ruzzo. Model-based clustering and data transformations for gene expression data. *Bioinformatics*, 17(10):977–987, 2001.

[36] K. R. Žalik and B. Žalik. Validity index for clusters of different sizes and densities. *Pattern Recognition Letters*, 32(2):221 – 234, 2011.