ICLR

Paper Code

Internal Evaluation of Density-Based Clusterings with Noise

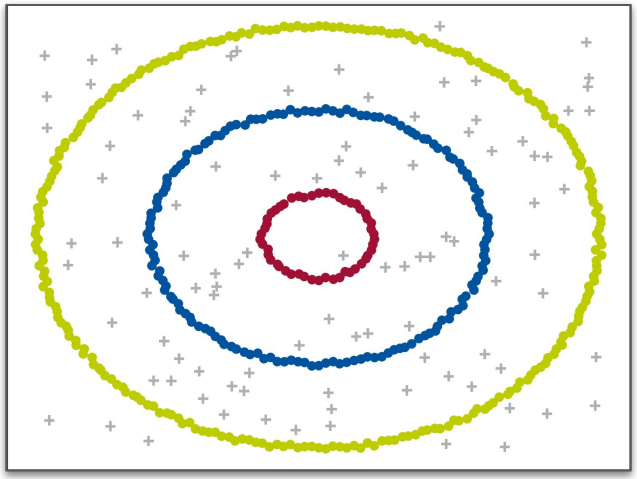
Anna Beer*[1,2], Lena Krieger*[3,4,5], Pascal Weber*[1,6,7], Martin Ritzert[8,9], Ira Assent[10], Claudia Plant[1,7]

[1]University of Vienna, Austria [2]Webster Vienna Private University, Austria [3]IAS-8, Forschungszentrum Jülich, Germany [4]LMU Munich, Germany [5]Munich Center for Machine Learning, Munich, Germany
[6]UniVie Doctoral School Computer Science, University of Vienna, Austria [7]Data Science @ Uni Vienna, University of Vienna, Austria
[8] University of Göttingen, Germany [9]Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI) Dresden/Leipzig, Leipzig University, Germany [10]Aarhus University, Denmark
*Equal Contribution.

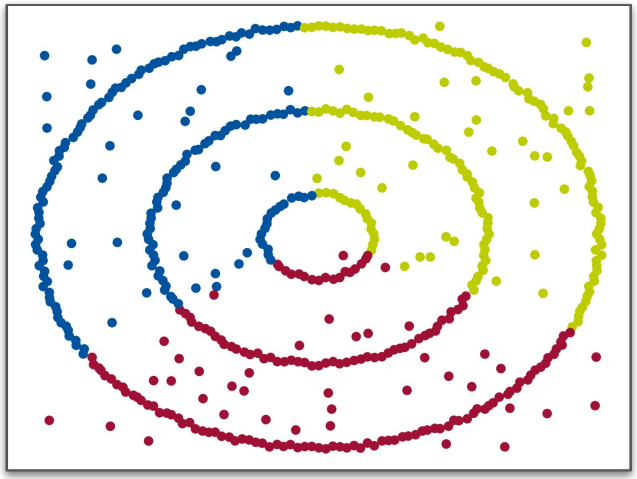
universität wien · Webster University · JÜLICH Forschungszentrum · LMU Ludwig-Maximilians-Universität München · mcml Munich Center for Machine Learning · Universität Leipzig · Aarhus University

Internal Evaluation of Clustering

- Great if ground truth is not known.
- Usually, a centroid-based clustering model is assumed.
- Arbitrary shapes and noise points are common, though.

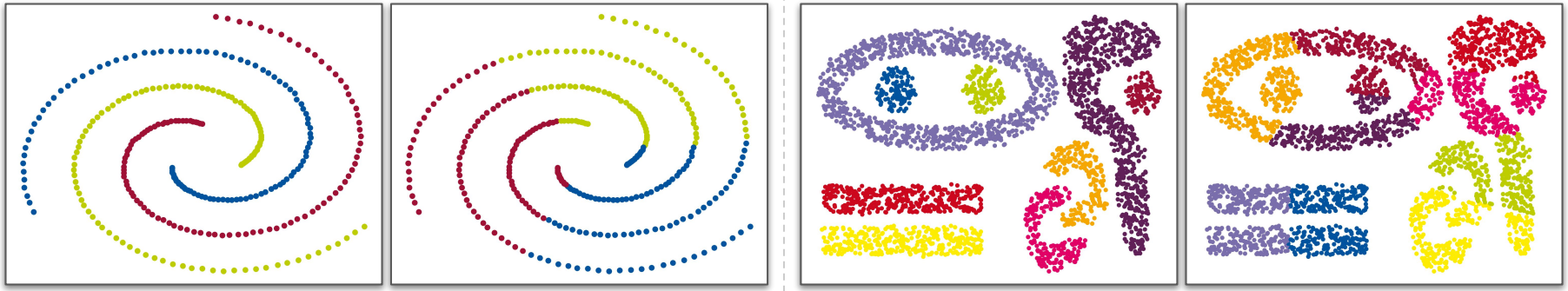


Low (bad) Silhouette Coefficient



High (good) Silhouette Coefficient

Quite some CVIs handle “arbitrary shapes”



3-spiral
DBSCAN

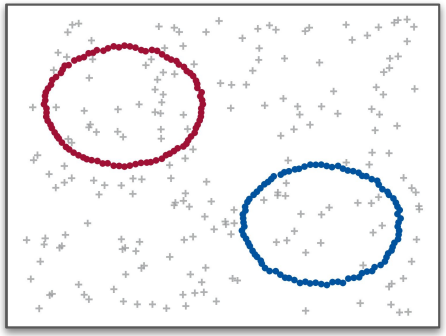
3-spiral
k-Means

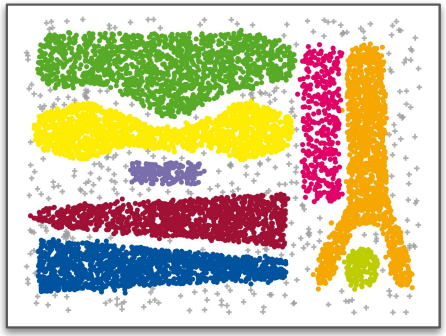
complex9
DBSCAN

complex9
k-Means

E.g.: DBCV, DCSI, LCCV, VIASCKDE, CVDD, CDbw, CVNN

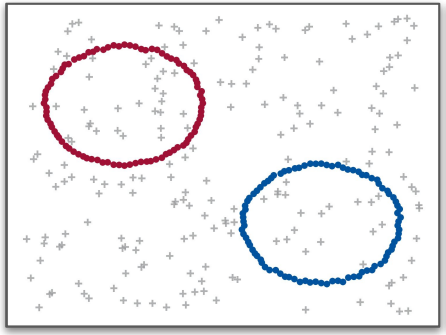
…but so far, only DBCV [1] is defined for clusterings with noise labels





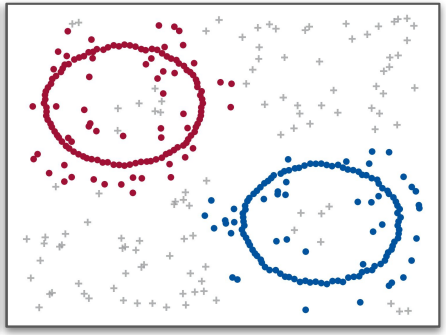
[1] Davoud Moulavi, Pablo A Jaskowiak, Ricardo J. G. B. Campello, Arthur Zimek, and Jörg Sander. Density-based clustering validation. In SIAM International Conference on Data Mining (SDM), pp. 839–847, 2014.

However, DBCV only uses the *number* of noise labels instead of evaluating their *quality*



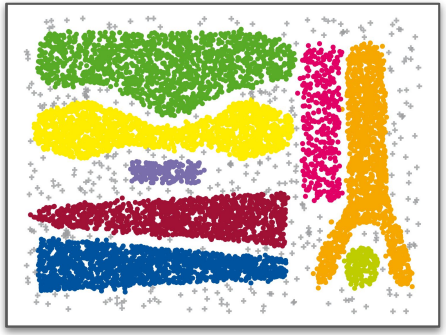
optimal

DBCV: 0.46



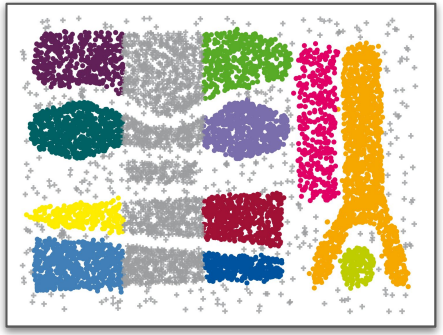
suboptimal

DBCV: 0.63



optimal

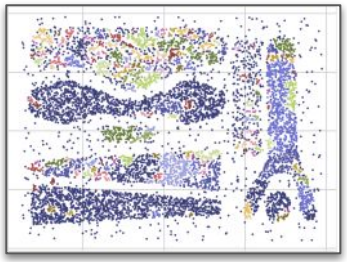
DBCV: -0.05



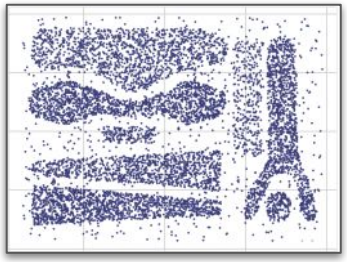
very bad

DBCV:0.17

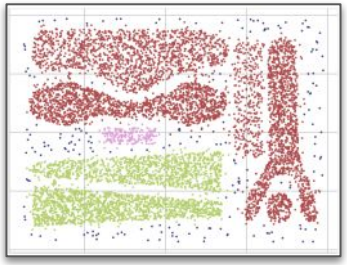
Major problem: DBCV is not deterministic.



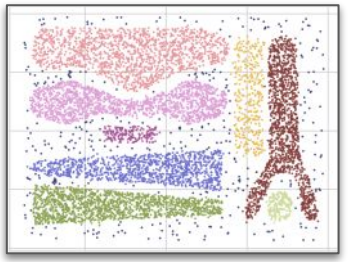
(a) 173 clusters + noise



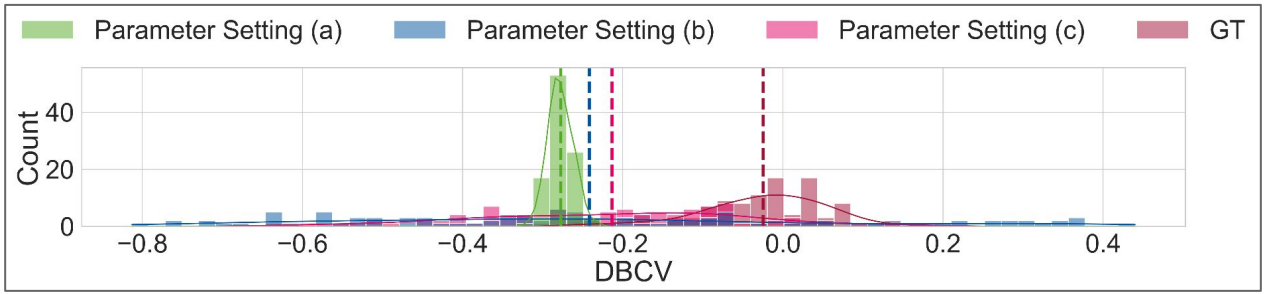
(b) 2 clusters



(c) 3 clusters + noise



GT: 8 clusters + noise



Why?

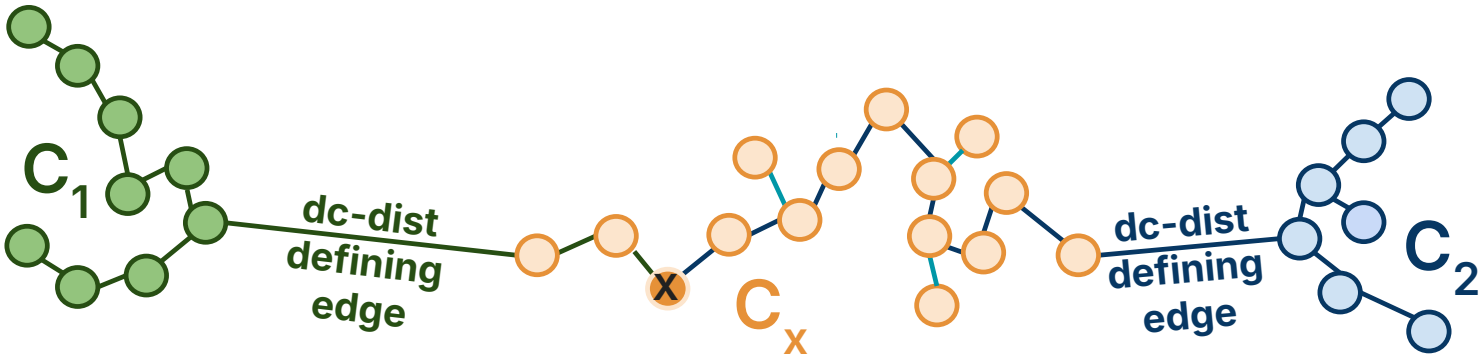
Internally, a MST is built (on top of the mutual reachability distance known from DBSCAN). The leaf nodes of this MST are removed for more robustness. But the MST is not unique.


Thus, we developed
DISCO
the first Density-based Internal Score for Clusterings with nOise

DISCO is a pointwise score based on density-connectivity, Silhouette Coefficient, and Noise Evaluation

$$\rho(x) = \begin{cases} \rho_{cluster}(x) & \text{if } x \in C_i \text{ for any } i \in [1, \ldots, k] \\ \rho_{noise}(x) & \text{if } x \in N \end{cases}$$

DISCO evaluates arbitrarily shaped clusters using the dc-distance [2]



The *density-connectivity distance* [2] is based on the mutual reachability distance d_m:

$$d_m(x, y) = \max\left(\kappa(x), \kappa(y), d_{eucl}(x, y)\right)$$

The **min-max path** on the graph given by d_m indicates the density-connectivity:

$$d_{dc}(x, y) = \max_{e \in p(x,y)} |e| \quad \text{if } x \neq y, \text{ else } 0$$

[2] Anna Beer, Andrew Draganov, Ellen Hohma, Philipp Jahn, Christian MM Frey, and Ira Assent. "Connecting the Dots--Density-Connectivity Distance unifies DBSCAN, k-Center and Spectral Clustering." In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 80-92. 2023.

…and inserts it into the **Silhouette Coefficient**

$$\rho_{cluster}(x) = \min_{C_i \neq \hat{C}_x} \frac{\widetilde{d_{dc}}(x, C_i) - \widetilde{d_{dc}}(x, \hat{C}_x)}{\max(\widetilde{d_{dc}}(x, C_i), \widetilde{d_{dc}}(x, \hat{C}_x))}$$

where $\widetilde{d_{dc}}(x, C_i) = \text{avg}_{y \in C_i} \, d_{dc}(x, y)$

DISCO evaluates noise labels

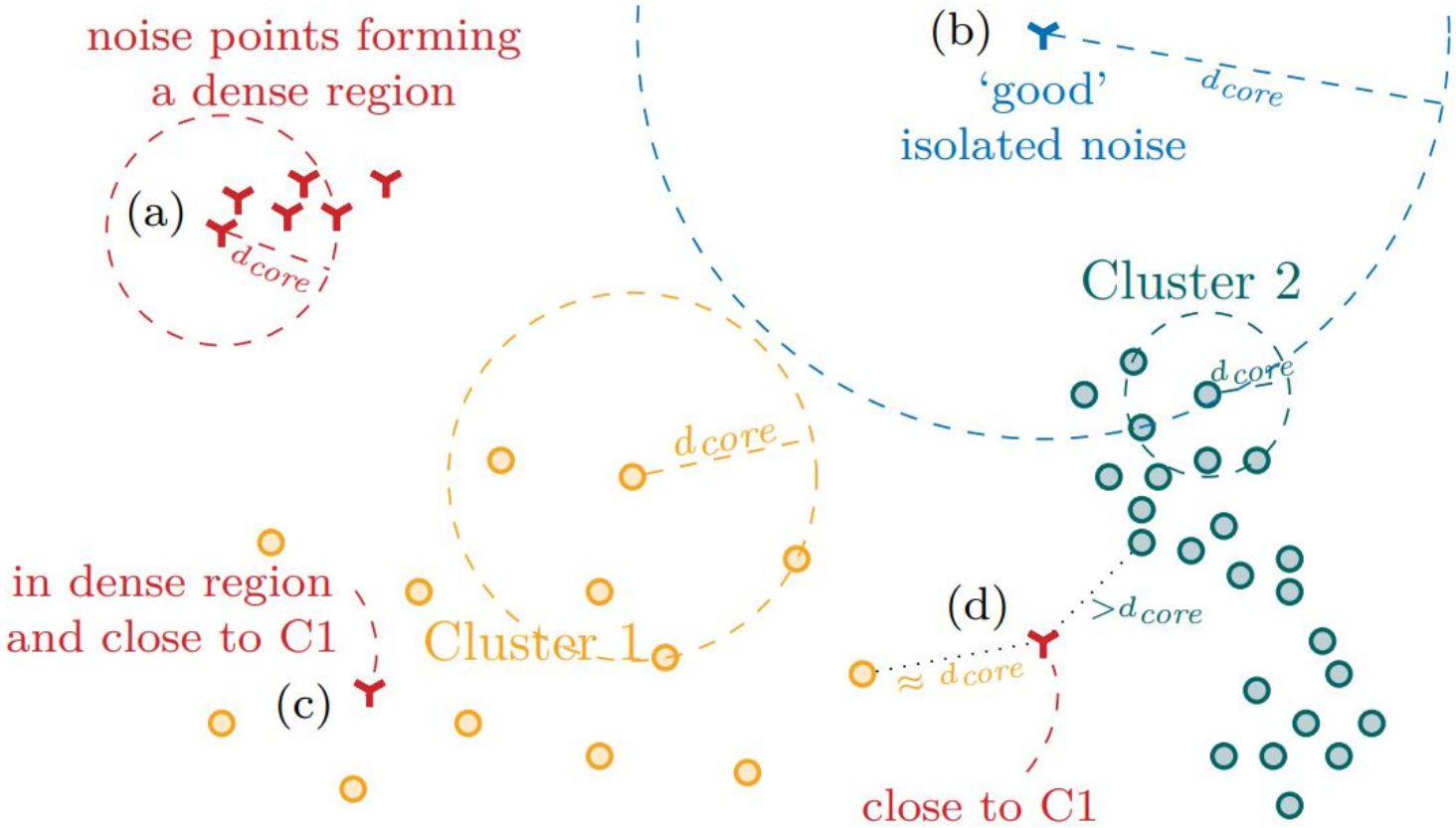
- based on the concepts from density-based clustering:
  Noise points are in **sparse** areas and **not density-connected** to any cluster

$$\rho_{noise}(x_n) = \min\left(\rho_{sparse}(x_n),\ \rho_{far}(x_n)\right)$$

$$\rho_{sparse}(x_n) = \min_{C_i \in \mathscr{C}} \frac{\kappa(x_n) - \kappa(C_i)}{\max\left(\kappa(x_n), \kappa(C_i)\right)}$$

$$\rho_{far}(x_n) = \min_{C_i \in \mathscr{C}} \frac{\min_{y \in C_i} d_{dc}(x_n, y) \ - \ \kappa(C_i)}{\max\left(\min_{y \in C_i} d_{dc}(x_n, y), \kappa(C_i)\right)}$$

DISCO evaluates noise labels



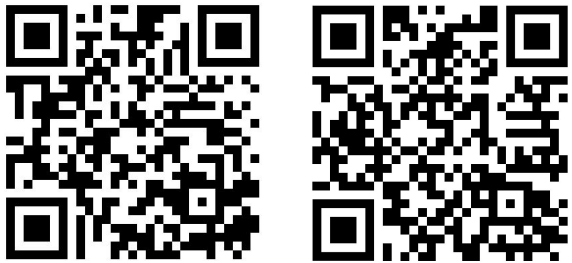
DISCO is defined on any clustering!

Clusterings with only one cluster, singleton clusters, only noise labels, or duplicate points lead to problems for other CVIs

Dataset	DISCO ↑	DBCV ↑	DCSI ↑	MMJ-SC ↑	LCCV ↑	VIAS. ↑	CVDD ↑	CDbw ↑	CVNN ↓	Silh. ↑	S_Dbw ↓
three_spiral	**89.16**	–	–	89.01	85.82	–	–	31.60	–	−2.43	42.38
aggregation	80.95	–	–	75.30	**92.41**	–	–	86.48	–	89.95	−81.13
chainlink	92.78	**99.71**	72.27	92.49	90.07	51.84	99.67	76.93	−36.53	26.99	26.50
cluto-t4-8k	44.43	77.56	**93.82**	62.68	76.12	57.02	18.37	–	−42.69	42.40	−53.84
cluto-t7-10k	48.66	83.87	**88.86**	61.64	32.42	50.38	−1.47	–	−39.42	13.49	−53.19
cluto-t8-8k	**91.35**	71.41	88.53	89.69	59.64	81.94	−1.53	–	−58.51	8.44	−68.10
complex8	95.71	90.53	90.09	**96.04**	90.17	87.15	47.60	48.43	−59.49	30.80	−71.59
complex9	56.35	59.63	**78.28**	61.20	75.16	68.58	19.25	66.13	−46.81	−1.52	−62.96
compound	86.08	–	–	87.05	**92.92**	–	–	67.59	–	62.12	−67.60
dartboard1	96.83	99.79	98.83	96.74	89.11	64.35	**99.95**	−53.39	−35.10	−20.07	−36.22
diamond9	98.99	87.13	**99.31**	99.27	93.52	98.99	67.20	13.71	−68.45	96.84	−87.33
smile1	**96.60**	–	96.40	96.58	94.62	–	–	68.53	–	79.20	−93.58
Synth_low	**98.13**	–	92.48	96.64	79.11	–	–	13.89	–	87.70	−85.88
Synth_high	**96.87**	–	95.52	96.30	72.72	–	–	56.44	–	88.93	−87.40
htru2	37.40	−41.94	−26.74	35.07	55.80	50.50	58.43	–	−37.98	**73.46**	−24.05
Pendigits	40.31	10.64	56.23	60.72	**79.03**	–	50.52	10.41	−43.92	78.22	−48.76
COIL20	95.79	93.44	94.17	**97.94**	93.13	–	63.99	21.84	−65.84	85.15	−90.29
cmu_faces	62.08	–	71.43	64.59	78.33	–	**80.75**	−2.84	−53.60	80.46	−55.85
Optdigits	91.07	50.57	83.32	**92.37**	90.14	–	65.70	12.44	−61.59	86.94	−70.67

DISCO!

–evaluates your density-based clustering with noise labels meaningfully–



Paper **Code**