**A study of clustering**

Eugen-Richard Ardelean1, Raluca Laura Portase1

1Department of Computer Science, Technical University of Cluj-Napoca, Cluj-Napoca, Romania

**\* Corresponding authors:** [ardeleaneugenrichard@gmail.com](mailto:ardeleaneugenrichard@gmail.com), [placehoder-raluca]

**ORCID Author IDs:**

Eugen-Richard Ardelean: 0000-0002-0098-4228

Raluca Laura Portase: 0000-0002-8985-4728

**Abstract: [**TBWritten]

**Keywords:** clustering, [TBAdded]

# Introduction

# Materials and Methods

## Clustering algorithms

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Name | Year |  | ?Complexity? | REF | CREF |  |
| K-Means | 1967 |  |  |  | SK |  |
| DBSCAN | 1996 |  |  |  | SK |  |
| Spectral Clustering | 2002 |  |  | <https://ai.stanford.edu/~ang/papers/nips01-spectral.pdf> | SK |  |
| MeanShift | 1995 |  |  |  | SK |  |
| Agglomerative Clustering | 1960-1970 |  |  |  | SK |  |
| OPTICS | 1999 |  |  | “OPTICS: ordering points to identify the clustering structure.” Ankerst, Mihael, Markus M. Breunig, Hans-Peter Kriegel, and Jörg Sander. In ACM Sigmod Record, vol. 28, no. 2, pp. 49-60. ACM, 1999. | SK |  |
| HDBSCAN | 2017 |  |  | L. McInnes and J. Healy, (2017). Accelerated Hierarchical Density Based Clustering. In: IEEE International Conference on Data Mining Workshops (ICDMW), 2017, pp. 33-42. Accelerated Hierarchical Density Based Clustering | SK |  |
| BIRCH | 1996 |  |  | Tian Zhang, Raghu Ramakrishnan, Maron Livny BIRCH: An efficient data clustering method for large databases. <https://www.cs.sfu.ca/CourseCentral/459/han/papers/zhang96.pdf> |  |  |
| DipInit | 2021 |  |  | Schelling, B., Bauer, L.G.M., Behzadi, S., Plant, C. (2021). Utilizing Structure-Rich Features to Improve Clustering. In: Hutter, F., Kersting, K., Lijffijt, J., Valera, I. (eds) Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2020. Lecture Notes in Computer Science(), vol 12457. Springer, Cham. https://doi.org/10.1007/978-3-030-67658-2\_6 | ClustPy |  |
| DipNSub | 2023 |  |  | Bauer, Lena GM, et al. "Extension of the Dip-test Repertoire-Efficient and Differentiable p-value Calculation for Clustering." Proceedings of the 2023 SIAM International Conference on Data Mining (SDM). Society for Industrial and Applied Mathematics, 2023. | ClustPy |  |
| Projected Dip-Means | 2018 |  |  | Theofilos Chamalis and Aristidis Likas. 2018. The Projected Dip-means Clustering Algorithm. In Proceedings of the 10th Hellenic Conference on Artificial Intelligence (SETN '18). Association for Computing Machinery, New York, NY, USA, Article 14, 1–7. https://doi.org/10.1145/3200947.3201008 | ClustPy |  |
|  |  |  |  |  |  |  |

## Clustering validation scores

Internal vs external – discussion from edging distance

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Type | Description | Range [worst, best] |
| ARI | External | Pair-by-pair comparison whether the points in the predicted cluster belong in the same true cluster | [-1, 1] |
| AMI | External | Mutual information based on entropy is used to calculate the agreement of true and predicted labels | [0, 1] |
| Purity | External | Cluster homogeneity as the majority class assignment. | [0, 1] |
| DBS | Internal | Ratio of the inter-cluster and intra-cluster sum of squared distances | (Inf, 0] |
| CHS | Internal | The average of a function that evaluates inter-cluster distances and the size of the cluster | [0, Inf) |
| SS | Internal | Cluster quality is evaluated as the balance between a cluster’s tightness and separation | [-1, 1] |

# Results

## Simple clustering

* Data table

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ID | Name | #Samples | #Features | #Clusters | Provenience |  |
|  | D1 | 1000 | 2 | 3 | Sk-learn |  |
|  | D2 | 1000 | 2 | 3 | Sk-learn |  |
|  | D3 | 1000 | 2 | 3 | Sk-learn |  |

@article{scikit-learn,

title={Scikit-learn: Machine Learning in {P}ython},

author={Pedregosa, F. and Varoquaux, G. and Gramfort, A. and Michel, V.

and Thirion, B. and Grisel, O. and Blondel, M. and Prettenhofer, P.

and Weiss, R. and Dubourg, V. and Vanderplas, J. and Passos, A. and

Cournapeau, D. and Brucher, M. and Perrot, M. and Duchesnay, E.},

journal={Journal of Machine Learning Research},

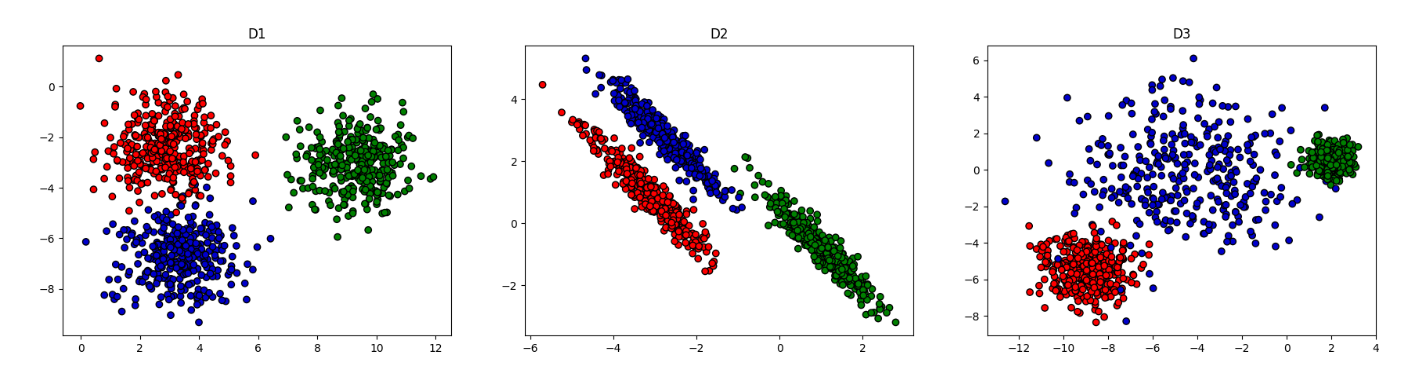
volume={12},

pages={2825--2830},

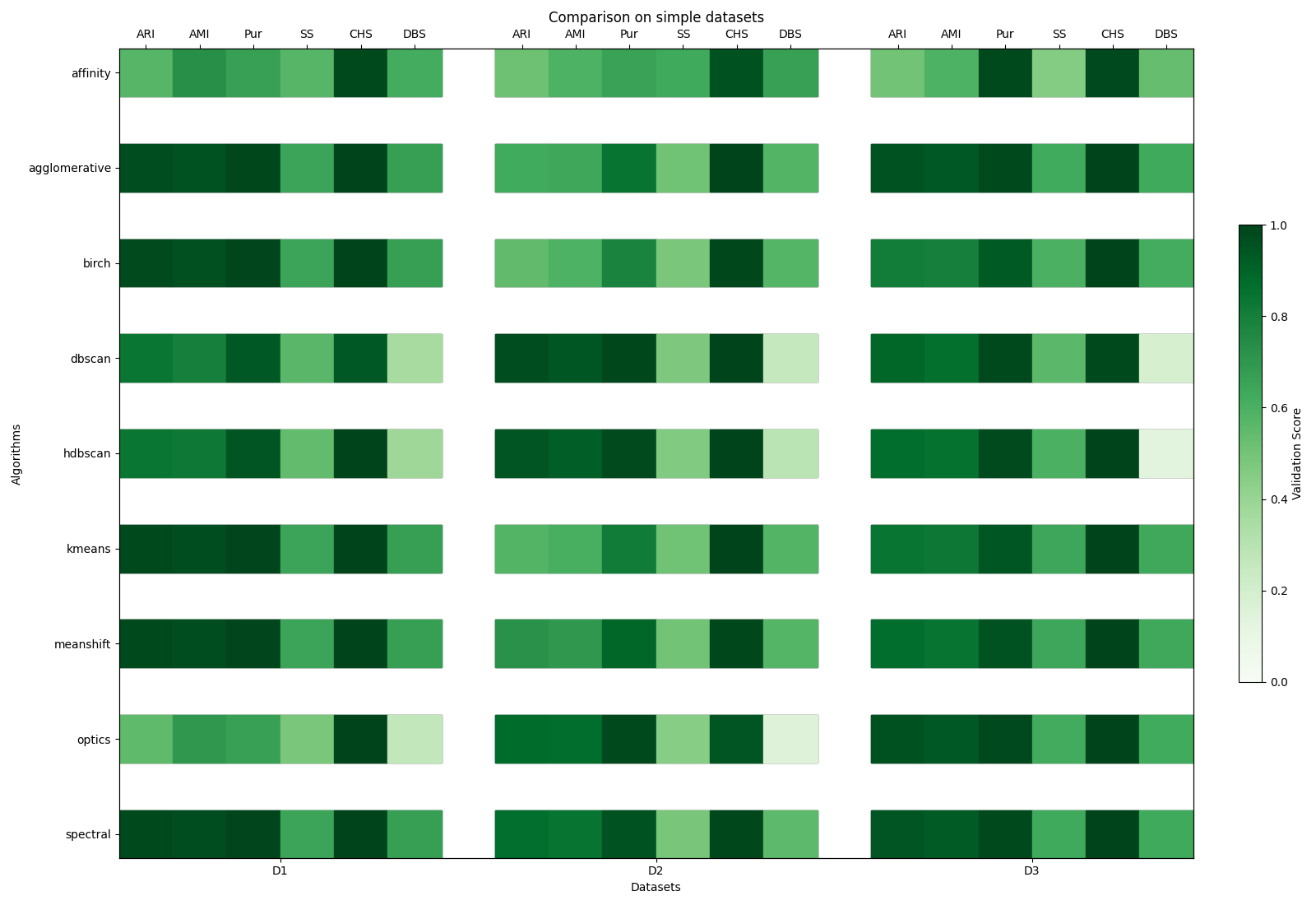
year={2011}

}

* Data plot



* Plot result – hierarchical



## Overlap

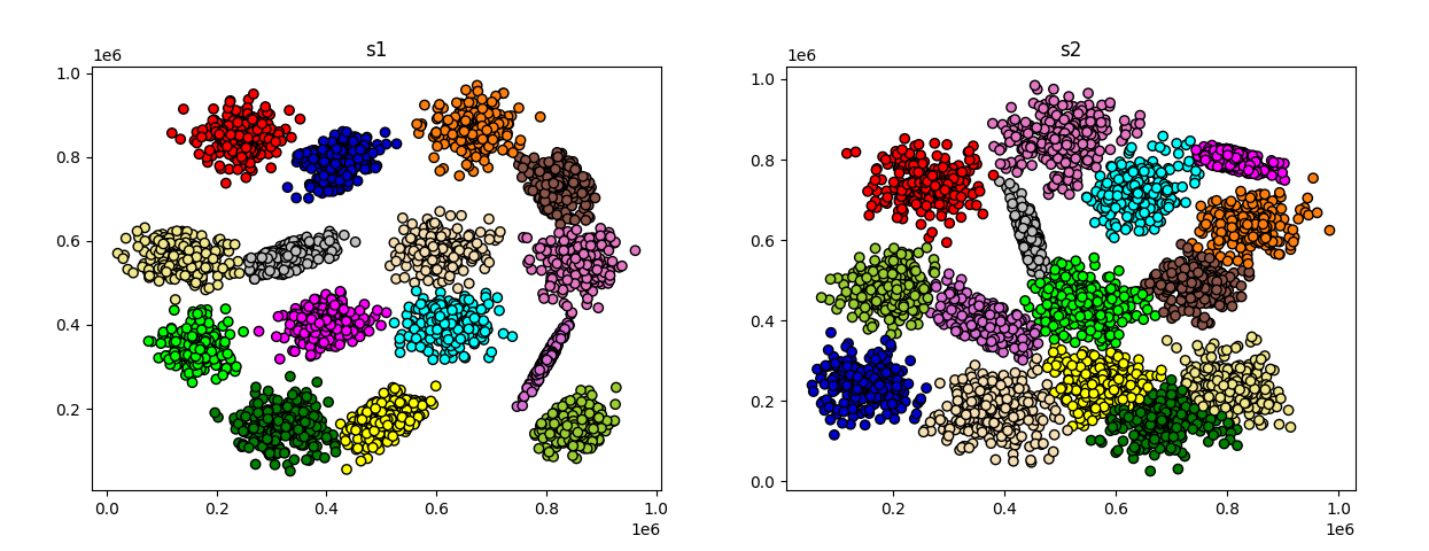
* Data table

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ID | Name | #Samples | #Features | #Clusters | Provenience |  |
|  | S1 | 5000 | 2 | 15 |  |  |
|  | S2 | 5000 | 2 | 15 |  |  |

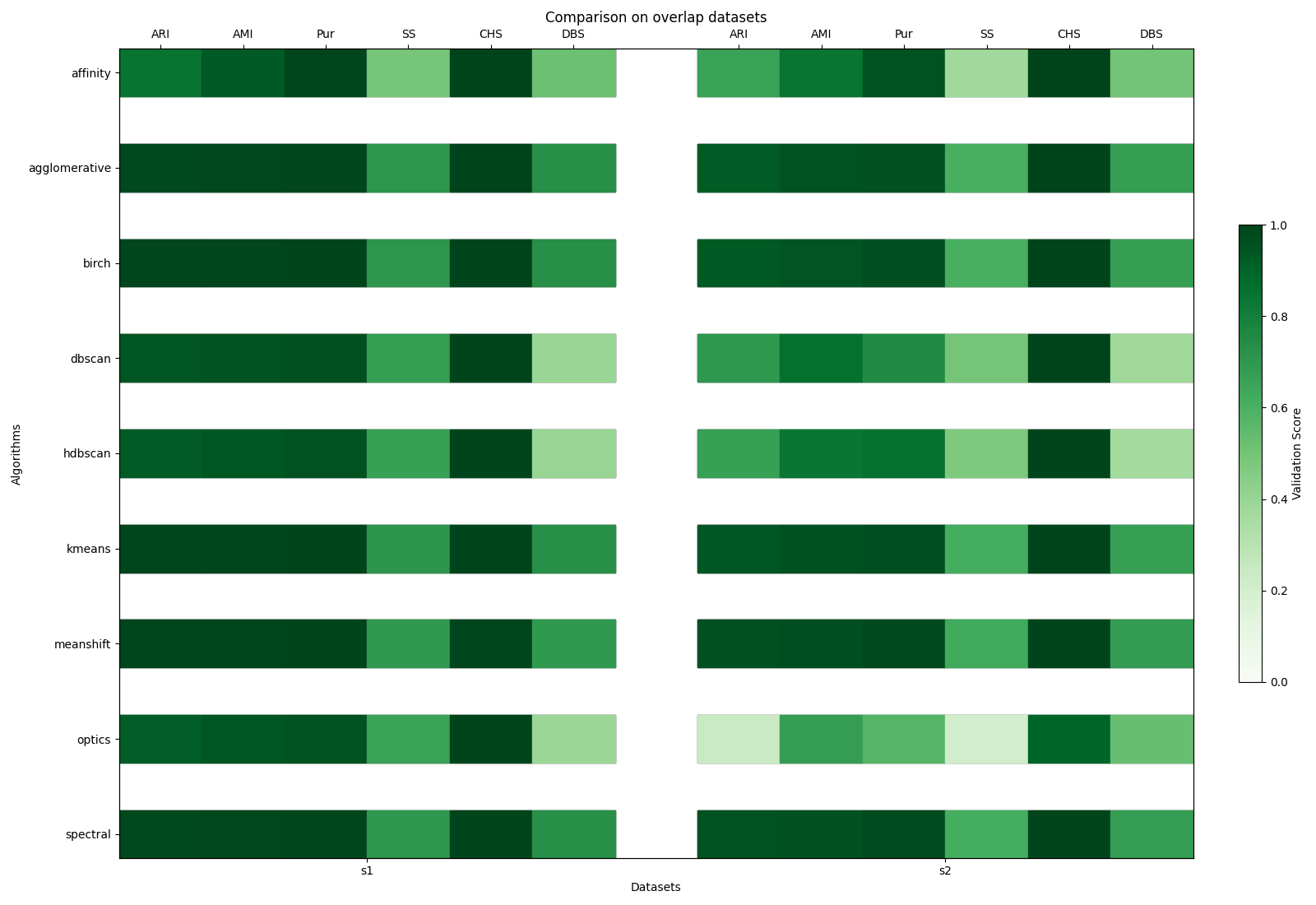
S1&S2

% P. Fränti and O. Virmajoki, "Iterative shrinking method for clustering problems", Pattern Recognition, 39 (5), 761-765, May 2006.

* Data plot



* Plot result – hierarchical

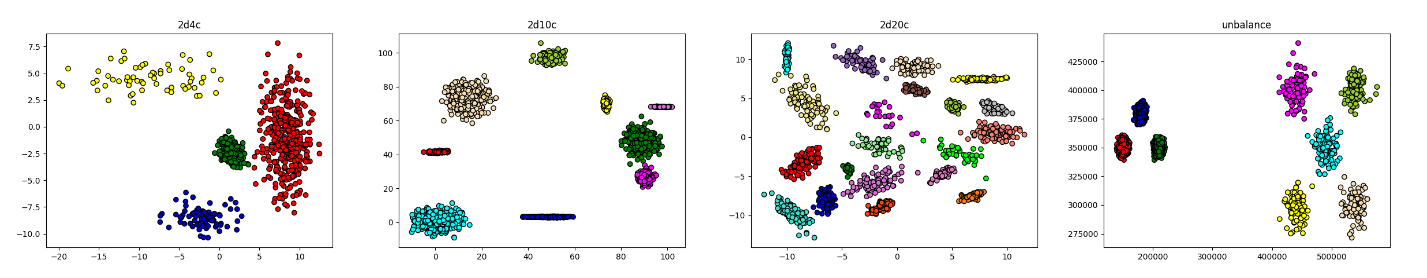


## Imbalance

* Data table

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ID | Name | #Samples | #Features | #Clusters | Provenience |  |
|  | 2d4c | 863 | 2 | 4 |  |  |
|  | 2d10c | 2990 | 2 | 9 |  |  |
|  | 2d20c | 1517 | 2 | 20 |  |  |
|  | unbalance | 6500 | 2 | 8 |  |  |

* Data plot

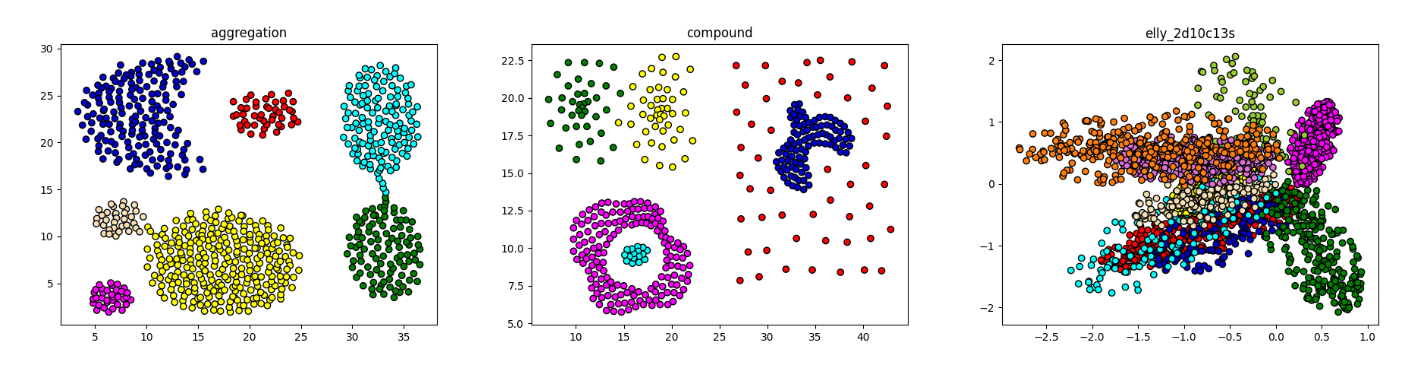


* Plot result – hierarchical

A chart of green squares

Description automatically generated

## Overlap and imbalance



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ID | Name | #Samples | #Features | #Clusters | Provenience |  |
|  | aggregation | 788 | 2 | 7 |  |  |
|  | compound | 399 | 2 | 6 |  |  |
|  | elly | 2796 | 2 | 10 |  |  |

AGGREGATION:

% Gionis, A., H. Mannila, and P. Tsaparas, Clustering aggregation. ACM Transactions on Knowledge Discovery from Data (TKDD), 2007. 1(1): p. 1-30.

% source: <http://cs.joensuu.fi/sipu/datasets/>

COMPOUND:

% Zahn, C.T., Graph-theoretical methods for detecting and describing gestalt clusters. IEEE Transactions on Computers, 1971. 100(1): p. 68-86.

% source: http://cs.joensuu.fi/sipu/datasets/

ELLY & 2dXc:

% Data from J. Handl

% <https://personalpages.manchester.ac.uk/staff/Julia.Handl/generators.html>

https://github.com/deric/clustering-benchmark

A chart of green squares

Description automatically generated

## Non-convexity

Internal metrics are not good, use external – discussion from edging distance

* Data table

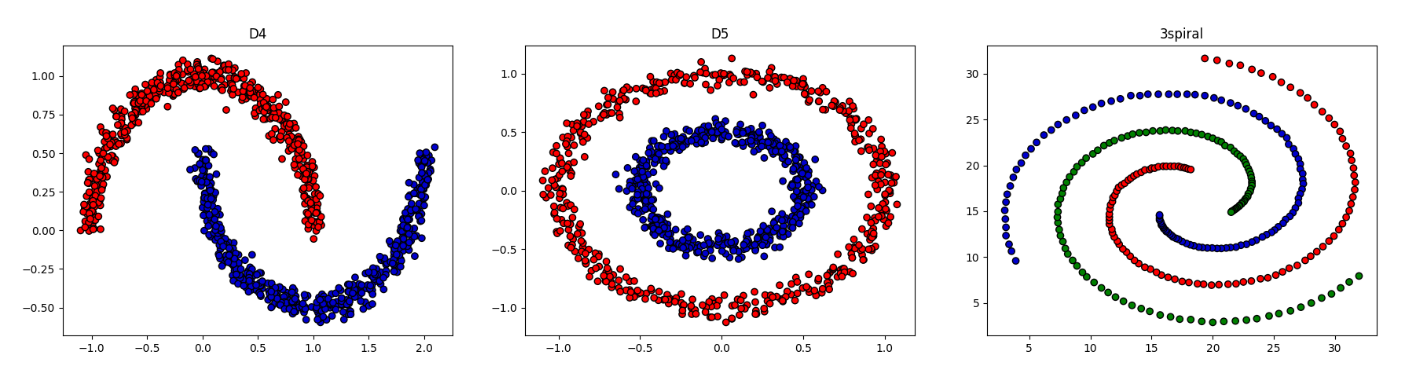
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ID | Name | #Samples | #Features | #Clusters | Provenience |  |
|  | D4 | 1000 | 2 | 2 | SK-LEARN |  |
|  | D5 | 1000 | 2 | 2 | SK-LEARN |  |
|  | 3spiral | 312 | 2 | 3 |  |  |

* Data plot

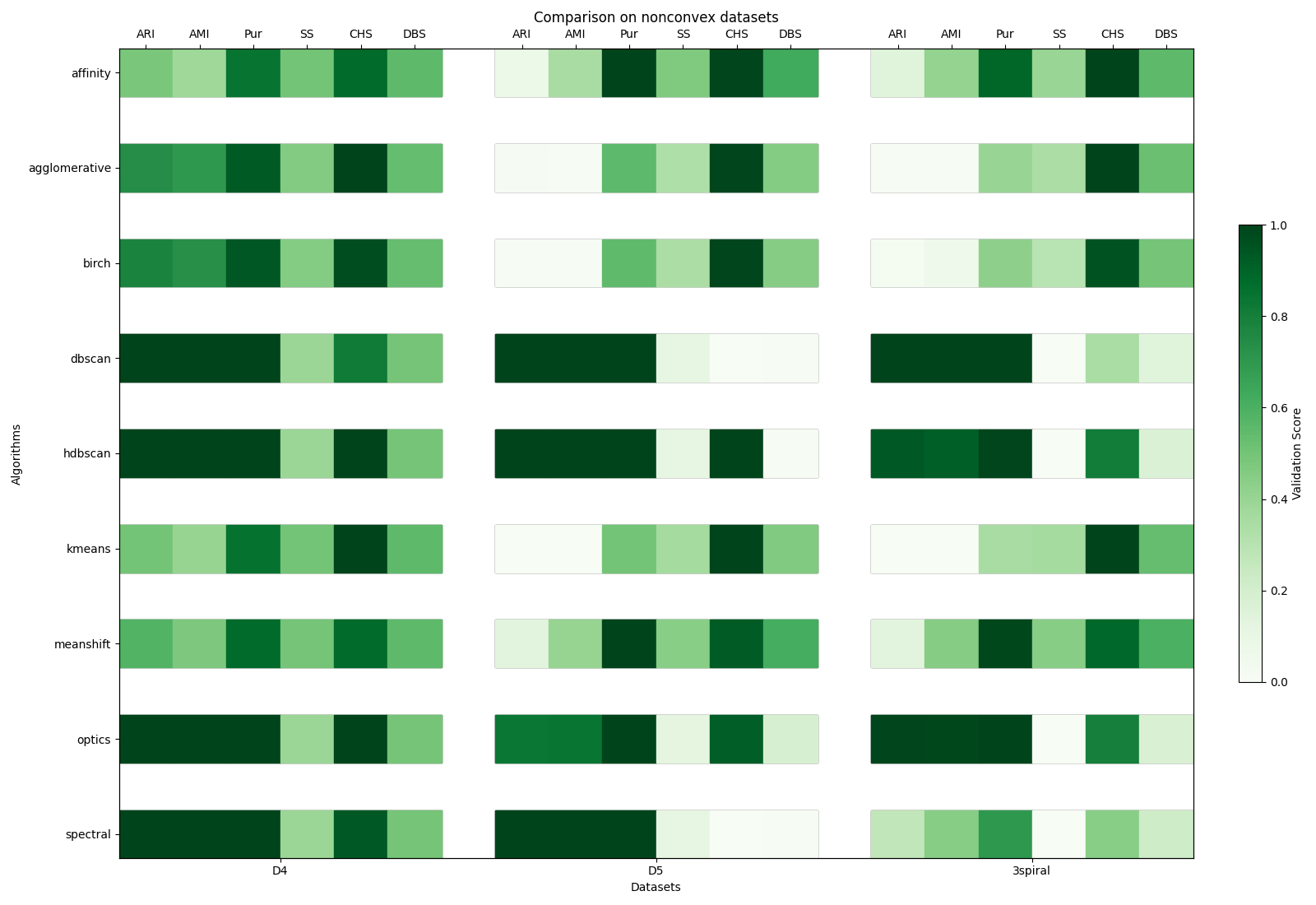
3SPIRAL:

% Chang, H. and D.Y. Yeung, Robust path-based spectral clustering. Pattern Recognition, 2008. 41(1): p. 191-203.

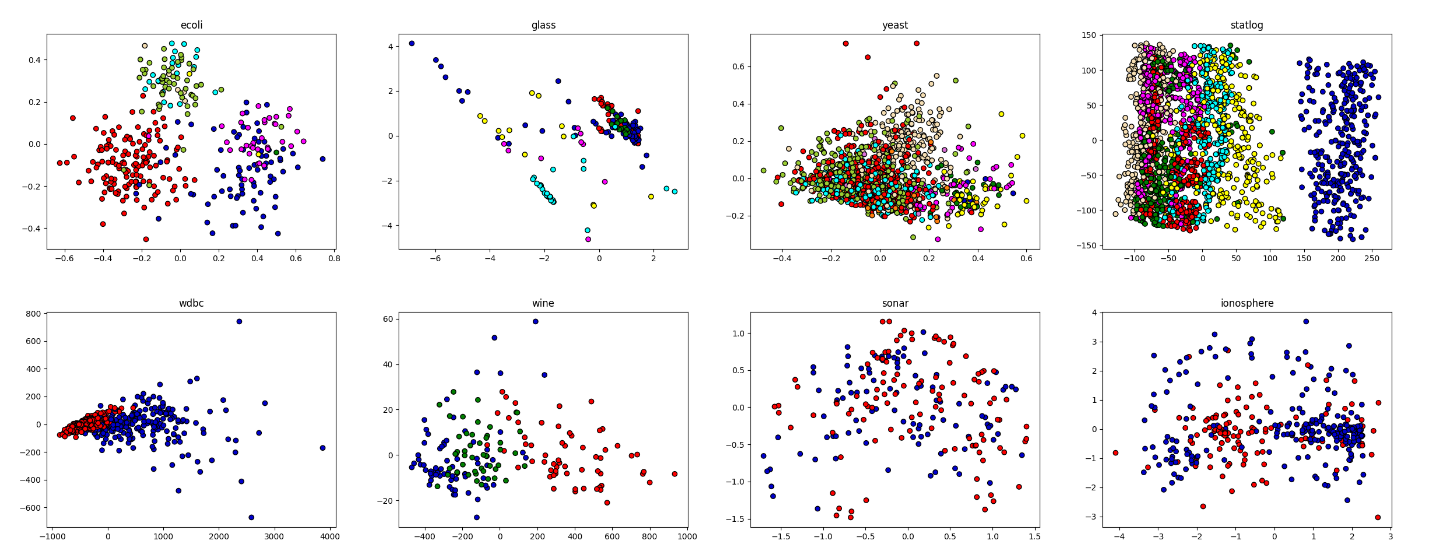
% source: http://cs.joensuu.fi/sipu/datasets/



* Plot result – hierarchical



## High dimensionality



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ID | Name | #Samples | #Features | #Clusters | Provenience |  |
|  | ecoli | 336 | 7 | 8 |  |  |
|  | glass | 214 | 9 | 6 |  |  |
|  | yeast | 1484 | 8 | 10 |  |  |
|  | statlog | 2310 | 19 | 7 |  |  |
|  | wdbc | 569 | 30 | 2 |  |  |
|  | wine | 178 | 13 | 3 |  |  |
|  | sonar | 208 | 60 | 2 |  |  |
|  | Ionosphere | 351 | 34 | 2 |  |  |

SONAR:

@misc{connectionist\_bench\_(sonar,\_mines\_vs.\_rocks)\_151,

author = {Sejnowski, Terry and Gorman, R.},

title = {{Connectionist Bench (Sonar, Mines vs. Rocks)}},

year = {1988},

howpublished = {UCI Machine Learning Repository},

note = {{DOI}: https://doi.org/10.24432/C5T01Q}

}

ECOLI:

@misc{ecoli\_39,

author = {Nakai, Kenta},

title = {{Ecoli}},

year = {1996},

howpublished = {UCI Machine Learning Repository},

note = {{DOI}: https://doi.org/10.24432/C5388M}

}

GLASS:

@misc{glass\_identification\_42,

author = {German, B.},

title = {{Glass Identification}},

year = {1987},

howpublished = {UCI Machine Learning Repository},

note = {{DOI}: https://doi.org/10.24432/C5WW2P}

}

YEAST:

@misc{yeast\_110,

author = {Nakai, Kenta},

title = {{Yeast}},

year = {1991},

howpublished = {UCI Machine Learning Repository},

note = {{DOI}: https://doi.org/10.24432/C5KG68}

}

STATLOG:

@misc{statlog\_(image\_segmentation)\_147,

title = {{Statlog (Image Segmentation)}},

year = {1990},

howpublished = {UCI Machine Learning Repository},

note = {{DOI}: https://doi.org/10.24432/C5P01G}

}

WINE:

@misc{wine\_109,

author = {Aeberhard, Stefan and Forina, M.},

title = {{Wine}},

year = {1992},

howpublished = {UCI Machine Learning Repository},

note = {{DOI}: https://doi.org/10.24432/C5PC7J}

}

WDBC:

@misc{breast\_cancer\_wisconsin\_(diagnostic)\_17,

author = {Wolberg, William, Mangasarian, Olvi, Street, Nick, and Street, W.},

title = {{Breast Cancer Wisconsin (Diagnostic)}},

year = {1993},

howpublished = {UCI Machine Learning Repository},

note = {{DOI}: https://doi.org/10.24432/C5DW2B}

}

Ionosphere:

@misc{ionosphere\_52,

author = {Sigillito, V., Wing, S., Hutton, L., and Baker, K.},

title = {{Ionosphere}},

year = {1989},

howpublished = {UCI Machine Learning Repository},

note = {{DOI}: https://doi.org/10.24432/C5W01B}

}

## Sample scalability

* Data table

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ID | Name | #Samples | #Features | #Clusters | Provenience |  |
| D1 |  |  |  |  |  |  |

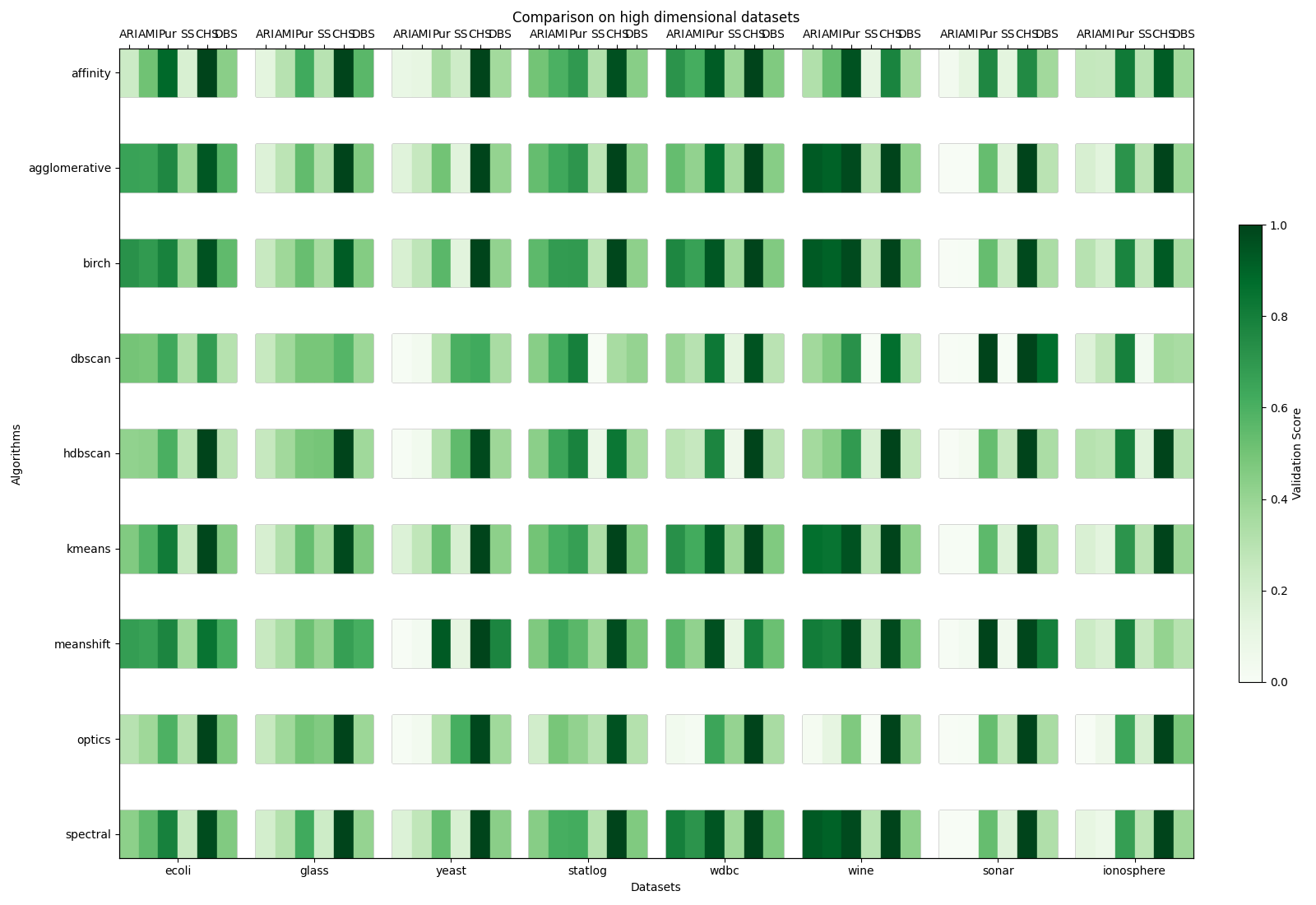
* Data plot
* Plot result – hierarchical

## Feature scalability

* Data table

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ID | Name | #Samples | #Features | #Clusters | Provenience |  |
| D1 |  |  |  |  |  |  |

* Data plot
* Plot result – hierarchical



# Discussion