**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> |  |  |
|  |  |  |  |  |  |  |

## 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] |

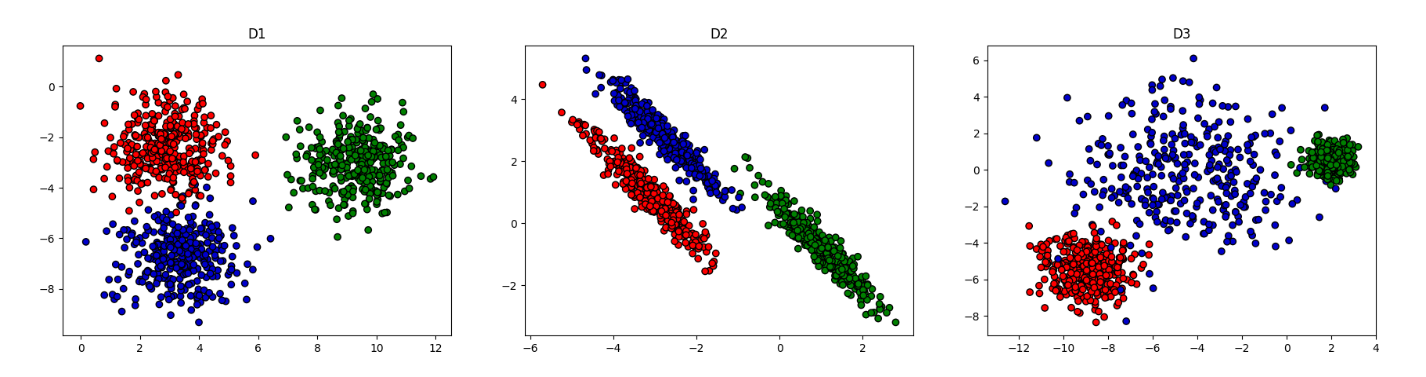
# Results

## Simple clustering

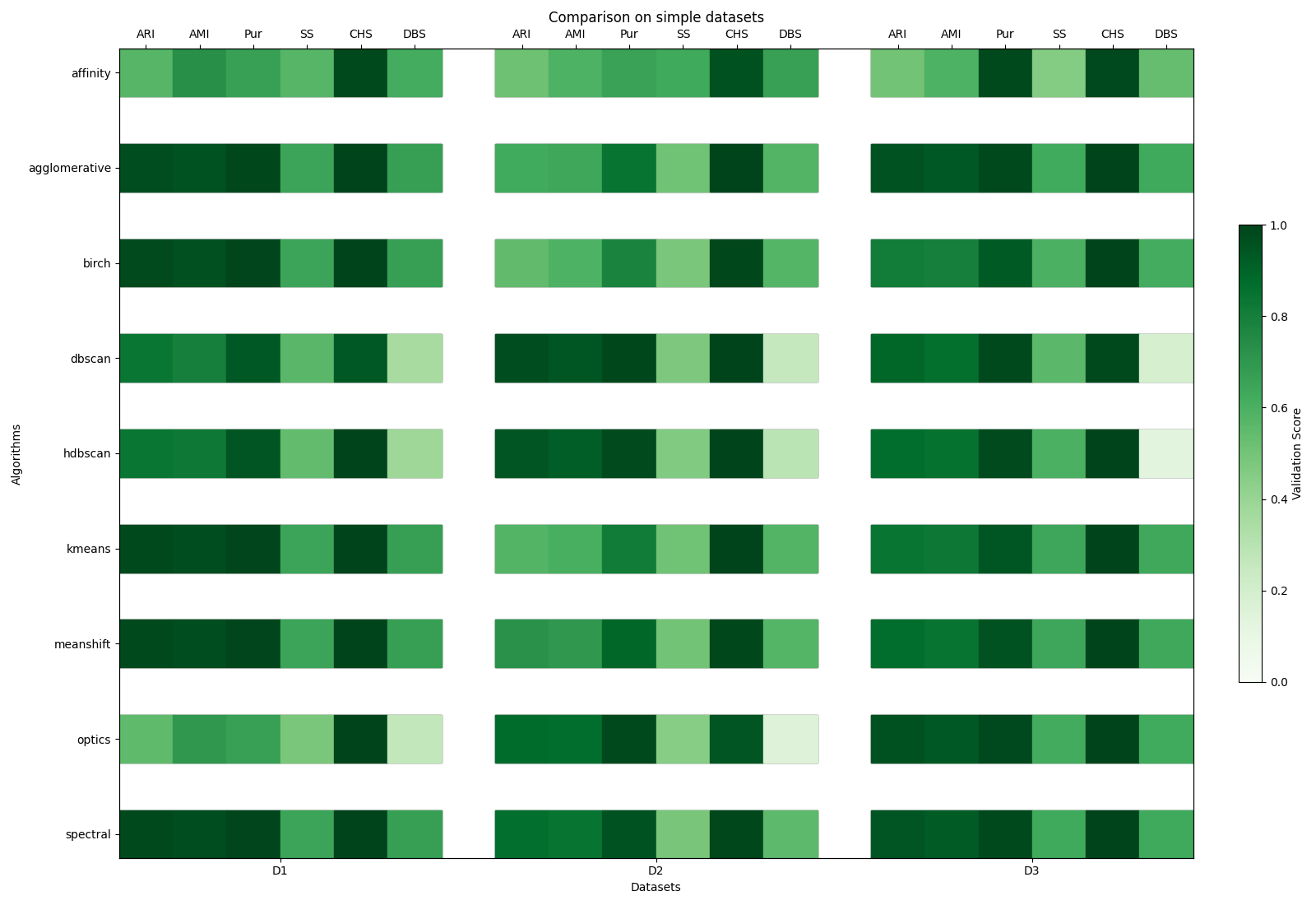
* Data table

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

* Data plot



* Plot result – hierarchical

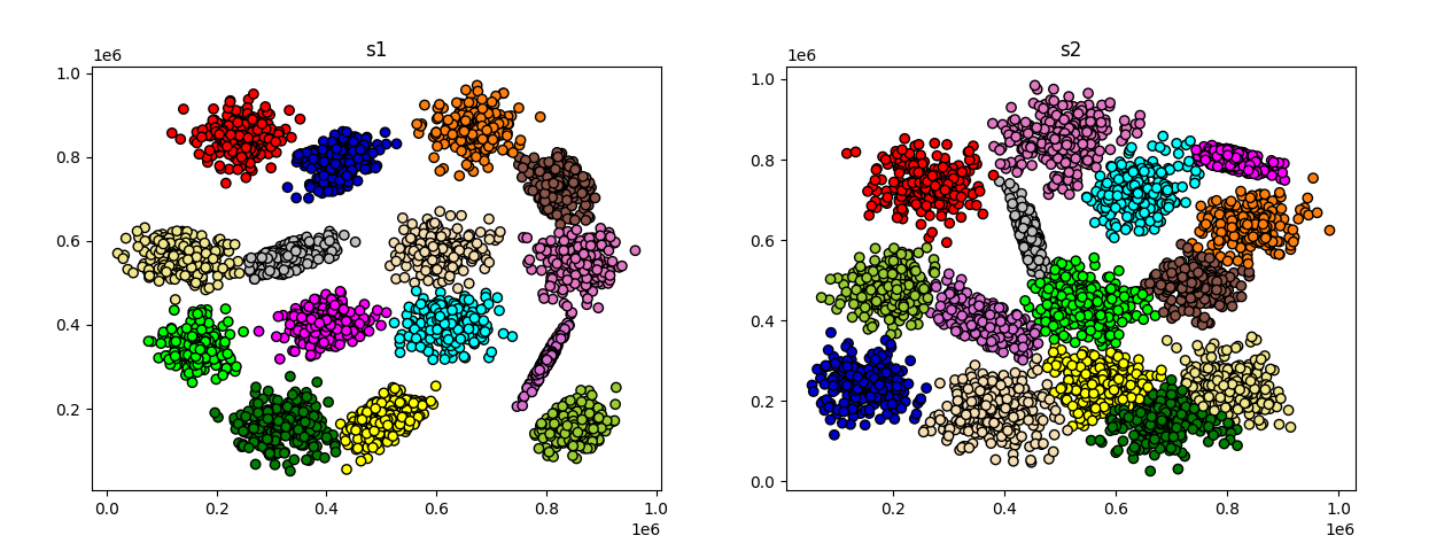


## Overlap

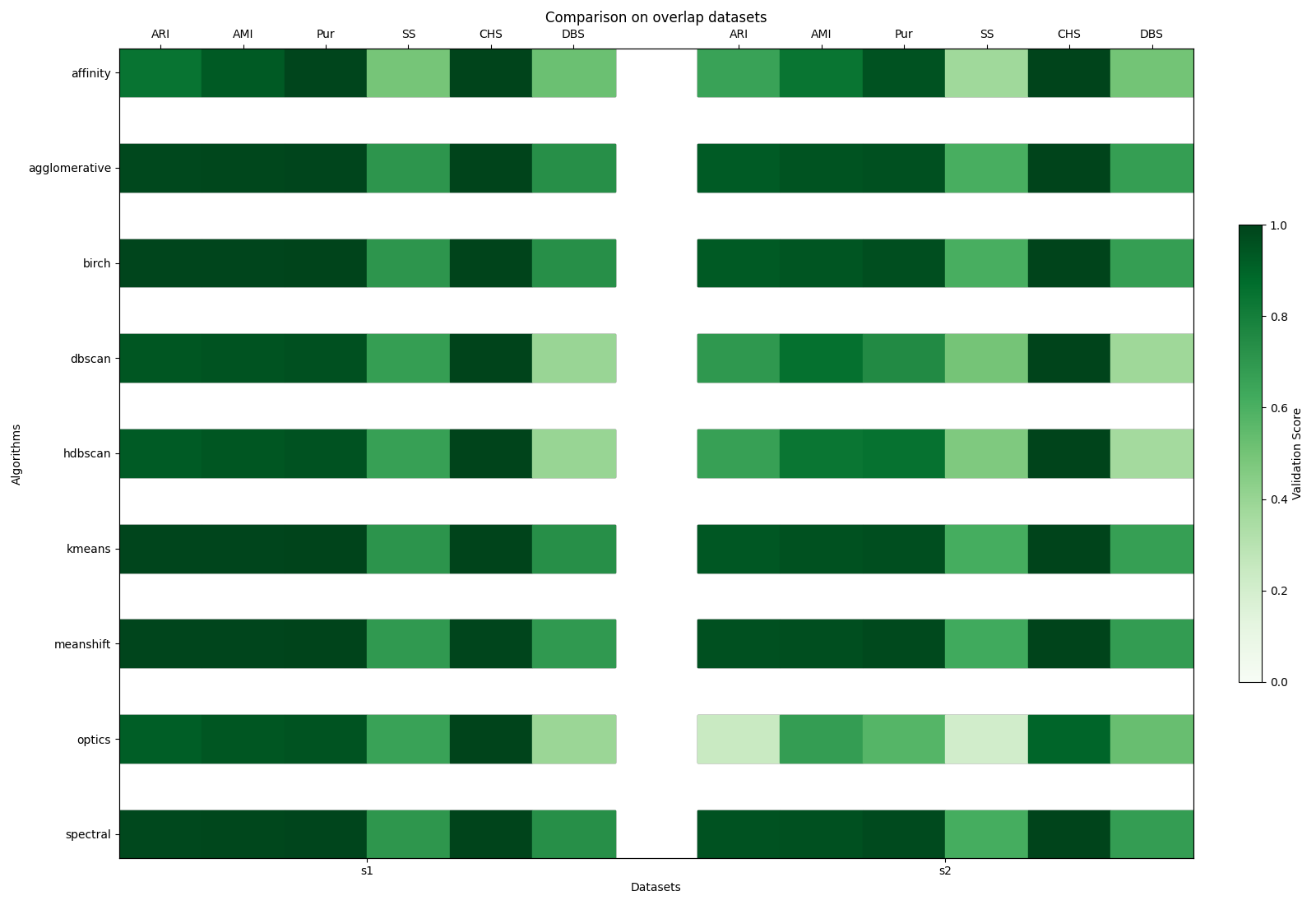
* Data table

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ID | Name | #Samples | #Features | #Clusters | Provenience |  |
| D1 | S1 | 5000 | 2 | 15 |  |  |
|  | S2 | 5000 | 2 | 15 |  |  |

* Data plot



* Plot result – hierarchical

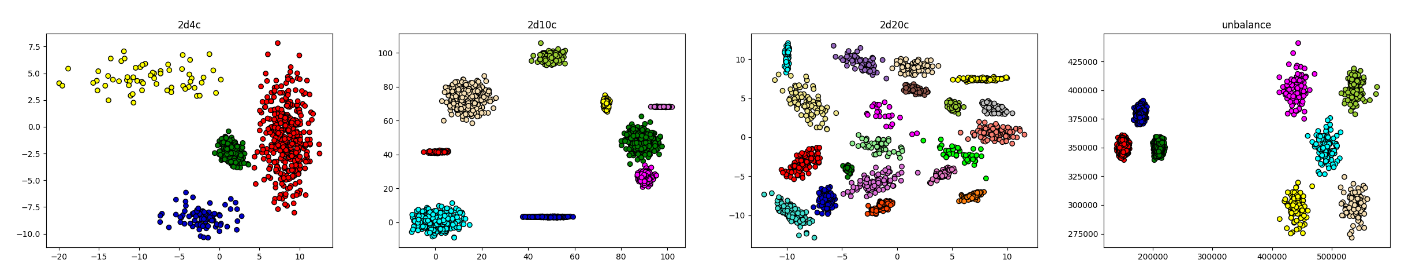


## Imbalance

* Data table

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ID | Name | #Samples | #Features | #Clusters | Provenience |  |
| D1 | 2d4c | 863 | 2 | 4 |  |  |
|  | 2d10c | 2990 | 2 | 9 |  |  |
|  | 2d20c | 1517 | 2 | 20 |  |  |
|  | unbalance | 6500 | 2 | 8 |  |  |

* Data plot

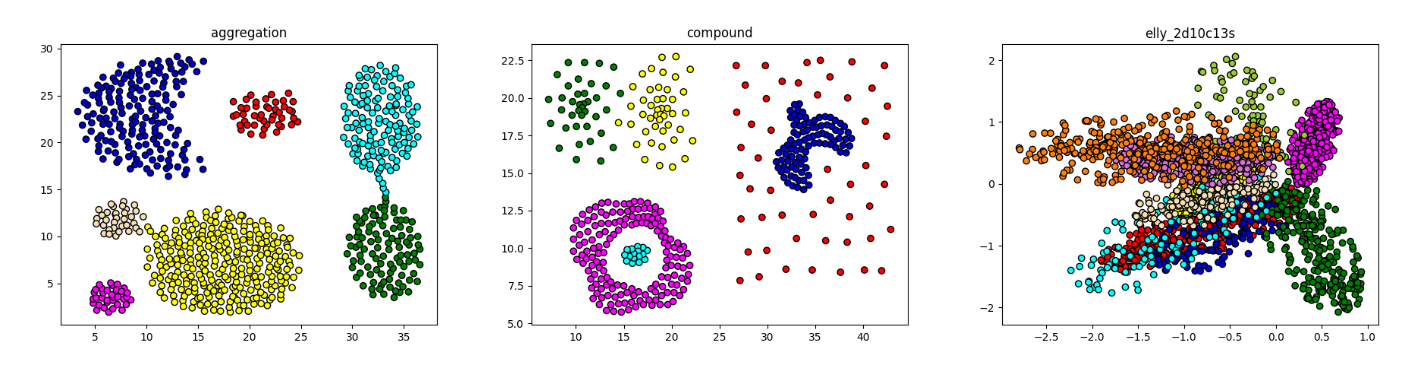


* Plot result – hierarchical

A chart of green squares

Description automatically generated

## Overlap and imbalance



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ID | Name | #Samples | #Features | #Clusters | Provenience |  |
| D1 | aggregation | 788 | 2 | 7 |  |  |
|  | compound | 399 | 2 | 6 |  |  |
|  | elly | 2796 | 2 | 10 |  |  |

A chart of green squares

Description automatically generated

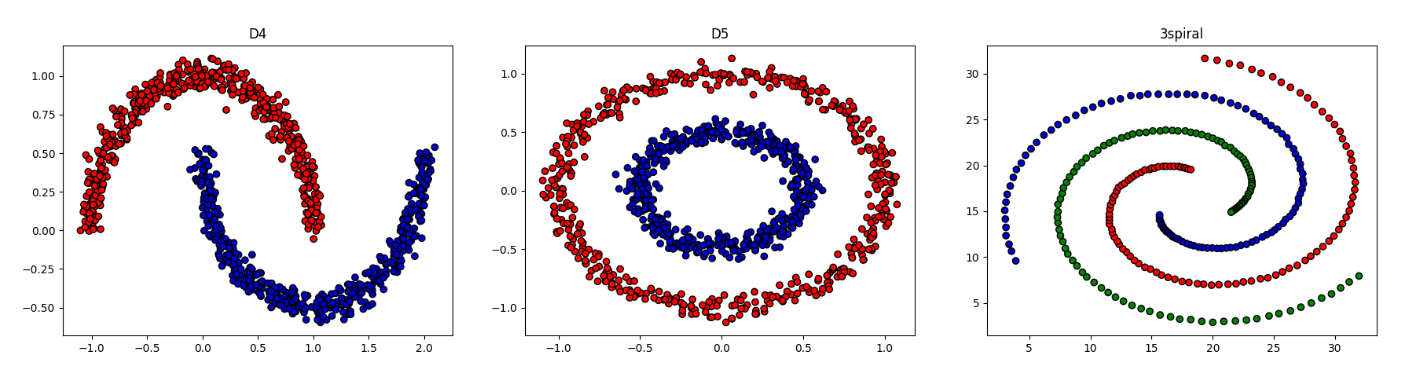
## Non-convexity

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

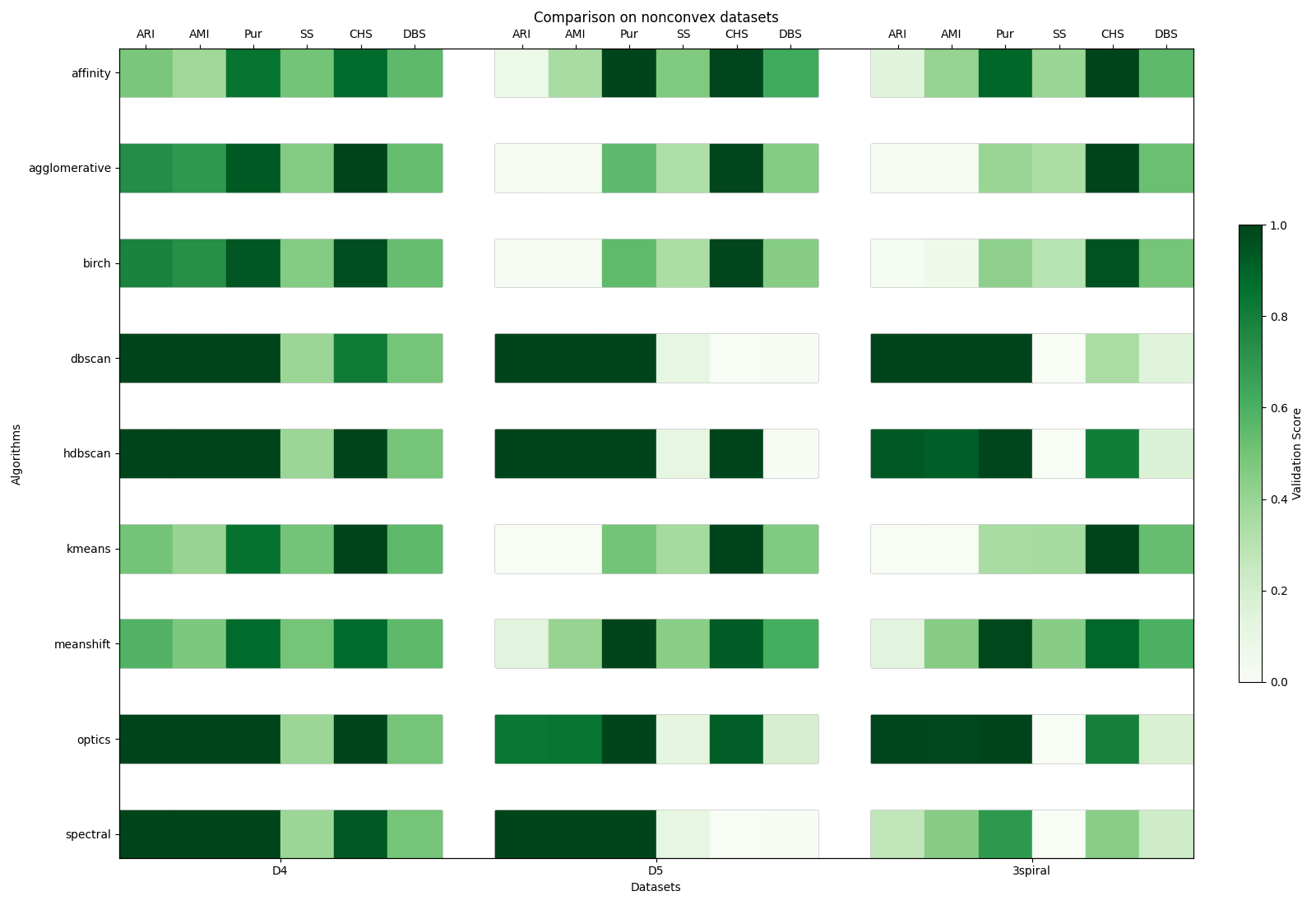
* Data table

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ID | Name | #Samples | #Features | #Clusters | Provenience |  |
| D1 | D4 | 1000 | 2 | 2 |  |  |
|  | D5 | 1000 | 2 | 2 |  |  |
|  | 3spiral | 312 | 2 | 3 |  |  |

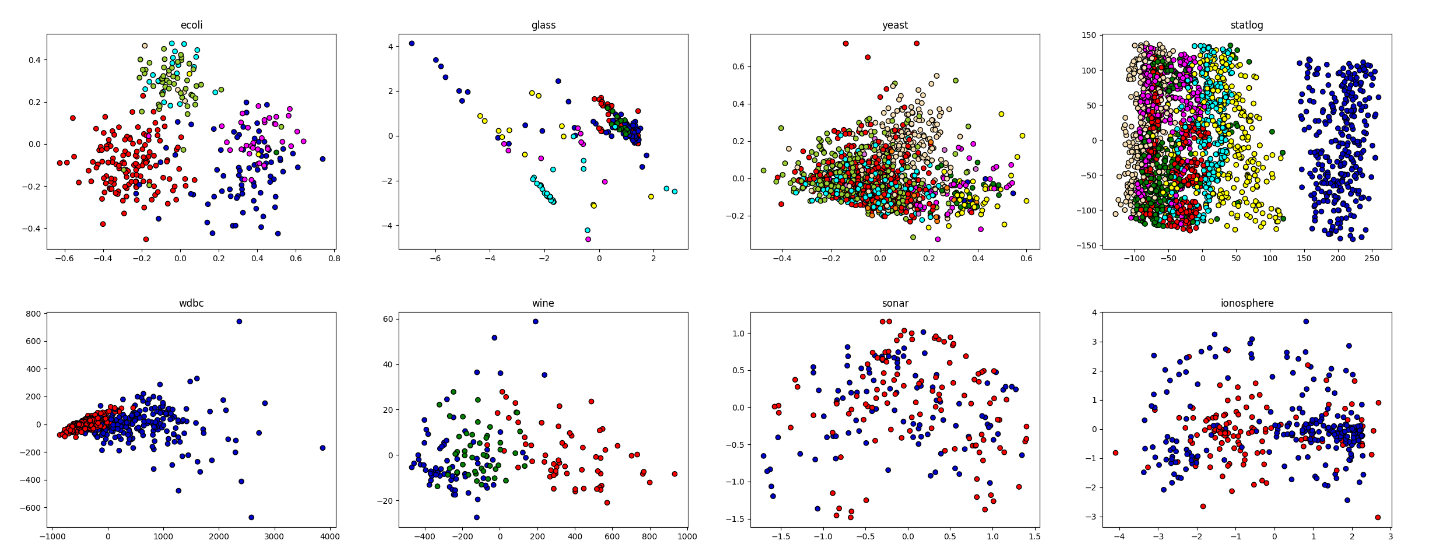
* Data plot



* 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 |  |  |

## Sample scalability

* Data table

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

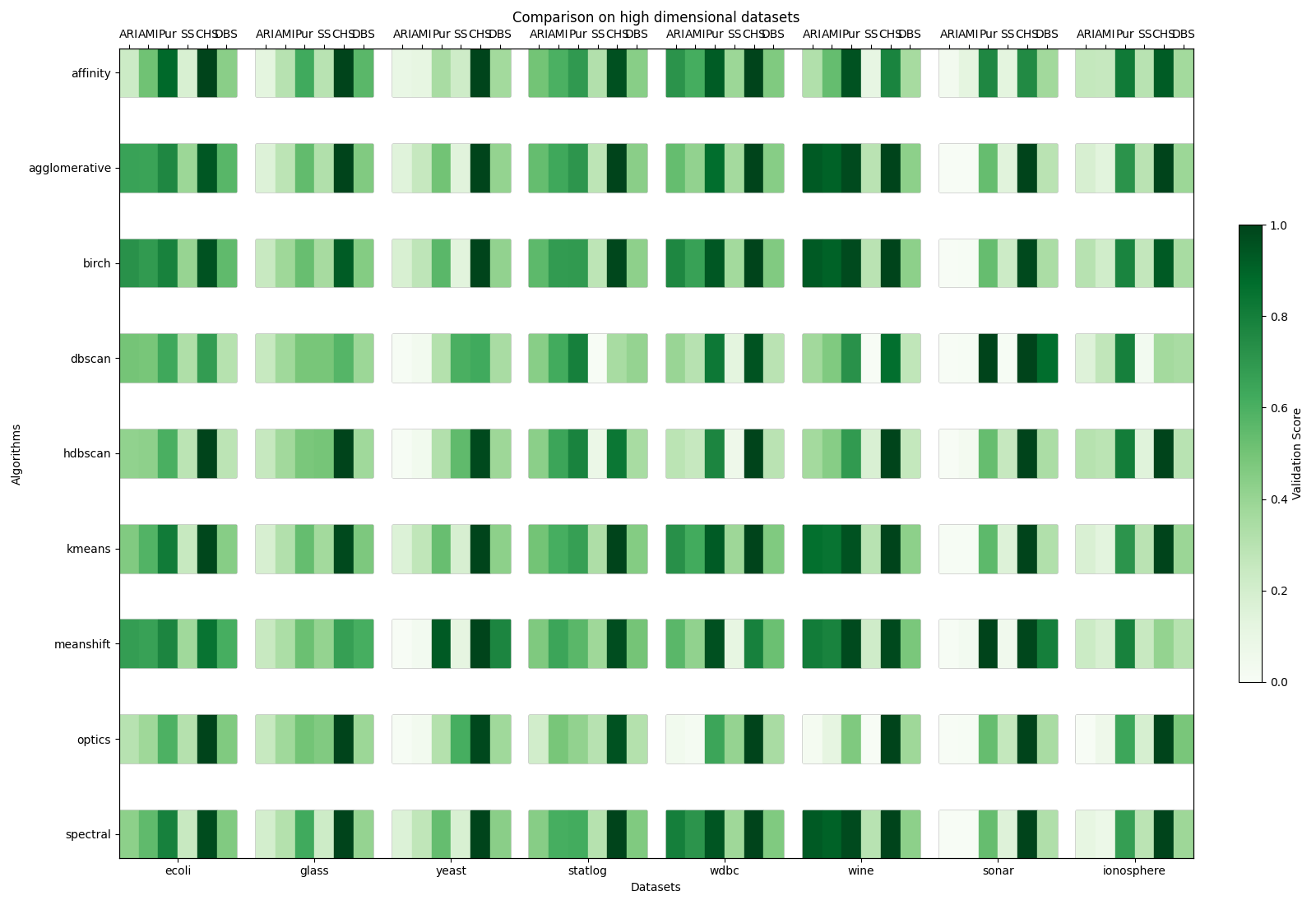
* Data plot
* Plot result – hierarchical

## Feature scalability

* Data table

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

* Data plot
* Plot result – hierarchical



# Discussion