# CLUSTERING

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

<https://github.com/gagolews/clustering-benchmarks>

<https://archive.ics.uci.edu/>

<https://github.com/collinleiber/ClustPy>

<https://github.com/annoviko/pyclustering>

<https://scikit-learn.org/stable/api/sklearn.cluster.html>

<https://scikit-learn.org/1.5/modules/clustering.html>

# K-MEANS

**Article:** "$t$-$k$-means: A Robust and Stable $k$-means Variant"

**Link:** [arXiv](https://arxiv.org/abs/1907.07442)

**Code:** [GitHub Repository](https://github.com/THUYimingLi/t-k-means) MATLAB

**Description:** This variant enhances robustness and stability by incorporating the Student's t-distribution into the K-Means framework.

**Article:** "Using MM Principles to Deal with Incomplete Data in K-means Clustering"

**Link:** [arXiv](https://arxiv.org/abs/2212.12379)

**Code:** [GitHub Repository](https://github.com/AliBeikmohammadi/MM-Optimization/blob/main/mini-project/MM%20K-means.ipynb) (some kid)

**Description:** This variation applies Majorization-Minimization (MM) principles to address incomplete data scenarios in K-Means clustering.

# DBSCAN

HDBSCAN

<https://scikit-learn.org/1.5/modules/clustering.html#hdbscan>

OPTICS

<https://scikit-learn.org/1.5/modules/clustering.html#optics>

**HDBSCAN: Hierarchical Density-Based Spatial Clustering of Applications with Noise**

* **Link:**<https://link.springer.com/article/10.1007/s10618-006-0057-x>
* **Code:**<https://github.com/scikit-learn-contrib/hdbscan>

**AMD-DBSCAN: An Adaptive Multi-density DBSCAN for datasets of extremely variable density**

* **Link:**<https://arxiv.org/abs/2210.08162>
* **Code:**<https://github.com/AlexandreWANG915/AMD-DBSCAN>

**KNN-DBSCAN: a DBSCAN in high dimensions**

* **Link:**<https://arxiv.org/abs/2009.04552>
* **Code:**<https://github.com/youguangchen/knn-dbscan> (not working)

**Automating DBSCAN via Deep Reinforcement Learning**

* **Link:**<https://arxiv.org/abs/2208.04537>
* **Code:**<https://github.com/RingBDStack/DRL-DBSCAN>

**Theoretically-Efficient and Practical Parallel DBSCAN**

* **Link:**<https://arxiv.org/abs/1912.06255>
* **Code:**<https://github.com/ParAlg/DBSCAN>

**MR-DBSCAN: A Scalable MapReduce-Based DBSCAN Algorithm for Heavily Skewed Data**

* **Link:**<https://ieeexplore.ieee.org/document/7055057>
* **Code:**<https://github.com/DSXiangLi/MR-DBSCAN> (not found)

**SNN-DBSCAN: Shared Nearest Neighbor Density-Based Clustering**

* **Link:**<https://link.springer.com/chapter/10.1007/3-540-44503-X_27>
* **Code:**<https://github.com/deric/clustering-benchmark/tree/master/src/algorithms/Shared%20Nearest%20Neighbor> (not found)

# CLUSTPY

## Partition-based Clustering

| **Algorithm** | **Publication** | **Published at** | **Original Code** | **Docs** |
| --- | --- | --- | --- | --- |
| DipInit (incl. DipExt) | [Utilizing Structure-Rich Features to Improve Clustering](https://link.springer.com/chapter/10.1007/978-3-030-67658-2_6) | ECML PKDD 2020 | [Link](https://figshare.com/articles/code/Utilizing_Structure-rich_Features_to_improve_Clustering/12063252/1) (R) | [Link](https://clustpy.readthedocs.io/en/latest/clustpy.partition.html#clustpy.partition.dipext.DipInit) |
| DipMeans | [Dip-means: an incremental clustering method for estimating the number of clusters](https://proceedings.neurips.cc/paper/2012/hash/a8240cb8235e9c493a0c30607586166c-Abstract.html) | NIPS 2012 | [Link](https://kalogeratos.com/psite/material/dip-means/) (Matlab) | [Link](https://clustpy.readthedocs.io/en/latest/clustpy.partition.html#clustpy.partition.dipmeans.DipMeans) |
| Dip'n'sub (incl. TailoredDip) | [Extension of the Dip-test Repertoire - Efficient and Differentiable p-value Calculation for Clustering](https://epubs.siam.org/doi/abs/10.1137/1.9781611977653.ch13) | SIAM SDM 2023 | [Link](https://figshare.com/articles/conference_contribution/Supplement_codes_and_data_for_the_paper_Extension_of_the_Dip-test_Repertoire_-_Efficient_and_Differentiable_p-value_Calculation_for_Clustering_/21916752) (Python) | [Link](https://clustpy.readthedocs.io/en/latest/clustpy.partition.html#clustpy.partition.dipnsub.DipNSub) |
| GapStatistic | [Estimating the number of clusters in a data set via the gap statistic](https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/1467-9868.00293) | RSS: Series B 2002 | - | [Link](https://clustpy.readthedocs.io/en/latest/clustpy.partition.html#clustpy.partition.gapstatistic.GapStatistic) |
| G-Means | [Learning the k in k-means](https://proceedings.neurips.cc/paper/2003/hash/234833147b97bb6aed53a8f4f1c7a7d8-Abstract.html) | NIPS 2003 | - | [Link](https://clustpy.readthedocs.io/en/latest/clustpy.partition.html#clustpy.partition.gmeans.GMeans) |
| LDA-K-Means | [Adaptive dimension reduction using discriminant analysis and K-means clustering](https://dl.acm.org/doi/abs/10.1145/1273496.1273562) | ICML 2007 | - | [Link](https://clustpy.readthedocs.io/en/latest/clustpy.partition.html#clustpy.partition.ldakmeans.LDAKmeans) |
| PG-Means | [PG-means: learning the number of clusters in data](https://proceedings.neurips.cc/paper/2006/hash/a9986cb066812f440bc2bb6e3c13696c-Abstract.html) | NIPS 2006 | - | [Link](https://clustpy.readthedocs.io/en/latest/clustpy.partition.html#clustpy.partition.pgmeans.PGMeans) |
| Projected Dip-Means | [The Projected Dip-means Clustering Algorithm](https://dl.acm.org/doi/abs/10.1145/3200947.3201008) | SETN 2018 | - | [Link](https://clustpy.readthedocs.io/en/latest/clustpy.partition.html#clustpy.partition.projected_dipmeans.ProjectedDipMeans) |
| SkinnyDip (incl. UniDip) | [Skinny-dip: Clustering in a Sea of Noise](https://dl.acm.org/doi/abs/10.1145/2939672.2939740) | KDD 2016 | [Link](https://github.com/samhelmholtz/skinny-dip) (R) | [Link](https://clustpy.readthedocs.io/en/latest/clustpy.partition.html#clustpy.partition.skinnydip.SkinnyDip) |
| SpecialK | [k Is the Magic Number—Inferring the Number of Clusters Through Nonparametric Concentration Inequalities](https://link.springer.com/chapter/10.1007/978-3-030-46150-8_16) | ECML PKDD 2019 | [Link](https://github.com/Sibylse/SpecialK) (Python) | [Link](https://clustpy.readthedocs.io/en/latest/clustpy.partition.html#clustpy.partition.specialk.SpecialK) |
| SubKmeans | [Towards an Optimal Subspace for K-Means](https://dl.acm.org/doi/abs/10.1145/3097983.3097989) | KDD 2017 | [Link](http://dmm.dbs.ifi.lmu.de/downloads/) (Scala) | [Link](https://clustpy.readthedocs.io/en/latest/clustpy.partition.html#clustpy.partition.subkmeans.SubKmeans) |
| X-Means | [X-means: Extending k-means with efficient estimation of the number of clusters](https://web.cs.dal.ca/~shepherd/courses/csci6403/clustering/xmeans.pdf) | ICML 2000 | - | [Link](https://clustpy.readthedocs.io/en/latest/clustpy.partition.html#clustpy.partition.xmeans.XMeans) |

## Density-based

| **Algorithm** | **Publication** | **Published at** | **Original Code** | **Docs** |
| --- | --- | --- | --- | --- |
| Multi Density DBSCAN | [Multi Density DBSCAN](https://link.springer.com/chapter/10.1007/978-3-642-23878-9_53) | IDEAL 2011 | - | [Link](https://clustpy.readthedocs.io/en/latest/clustpy.density.html#clustpy.density.multi_density_dbscan.MultiDensityDBSCAN) |

## Hierarchical

| **Algorithm** | **Publication** | **Published at** | **Original Code** | **Docs** |
| --- | --- | --- | --- | --- |
| DIANA | [Finding Groups in Data: An Introduction to Cluster Analysis](https://www.jstor.org/stable/2290430?origin=crossref) | JASA 1991 | - | [Link](https://clustpy.readthedocs.io/en/latest/clustpy.hierarchical.html#clustpy.hierarchical.diana.Diana) |

## Deep clustering

| **Algorithm** | **Publication** | **Published at** | **Original Code** | **Docs** |
| --- | --- | --- | --- | --- |
| ACe/DeC | [Details (Don't) Matter: Isolating Cluster Information in Deep Embedded Spaces](https://www.ijcai.org/proceedings/2021/389) | IJCAI 2021 | [Link](https://gitlab.cs.univie.ac.at/lukas/acedec_public) (Python + PyTorch) | [Link](https://clustpy.readthedocs.io/en/latest/clustpy.deep.html#clustpy.deep.enrc.ACeDeC) |
| AEC | [Auto-encoder based data clustering](https://link.springer.com/chapter/10.1007/978-3-642-41822-8_15) | CIARP 2013 | [Link](https://github.com/developfeng/DeepClustering/) (Matlab) | [Link](https://clustpy.readthedocs.io/en/latest/clustpy.deep.html#clustpy.deep.aec.AEC) |
| DCN | [Towards K-means-friendly spaces: simultaneous deep learning and clustering](https://dl.acm.org/doi/abs/10.5555/3305890.3306080) | ICML 2017 | [Link](https://github.com/boyangumn/DCN) (Python + Theano) | [Link](https://clustpy.readthedocs.io/en/latest/clustpy.deep.html#clustpy.deep.dcn.DCN) |
| DDC | [Deep density-based image clustering](https://www.sciencedirect.com/science/article/pii/S0950705120302112) | Knowledge-Based Systems 2020 | [Link](https://github.com/Yazhou-Ren/DDC/tree/master) (Python + Keras) | [Link](https://clustpy.readthedocs.io/en/latest/clustpy.deep.html#clustpy.deep.ddc_n2d.DDC) |
| DEC | [Unsupervised deep embedding for clustering analysis](https://dl.acm.org/doi/abs/10.5555/3045390.3045442) | ICML 2016 | [Link](https://github.com/piiswrong/dec) (Python + Caffe) | [Link](https://clustpy.readthedocs.io/en/latest/clustpy.deep.html#clustpy.deep.dec.DEC) |
| DeepECT | [Deep embedded cluster tree](https://ieeexplore.ieee.org/abstract/document/8970987) | ICDM 2019 | [Link](https://dmm.dbs.ifi.lmu.de/downloads) (Python + PyTorch) | [Link](https://clustpy.readthedocs.io/en/latest/clustpy.deep.html#clustpy.deep.deepect.DeepECT) |
| DipDECK | [Dip-based Deep Embedded Clustering with k-Estimation](https://dl.acm.org/doi/10.1145/3447548.3467316) | KDD 2021 | [Link](https://dmm.dbs.ifi.lmu.de/downloads) (Python + PyTorch) | [Link](https://clustpy.readthedocs.io/en/latest/clustpy.deep.html#clustpy.deep.dipdeck.DipDECK) |
| DipEncoder | [The DipEncoder: Enforcing Multimodality in Autoencoders](https://dl.acm.org/doi/10.1145/3534678.3539407) | KDD 2022 | [Link](https://dmm.dbs.ifi.lmu.de/downloads) (Python + PyTorch) | [Link](https://clustpy.readthedocs.io/en/latest/clustpy.deep.html#clustpy.deep.dipencoder.DipEncoder) |
| DKM | [Deep k-Means: Jointly clustering with k-Means and learning representations](https://www.sciencedirect.com/science/article/abs/pii/S0167865520302749) | Pattern Recognition Letters 2020 | [Link](https://github.com/MaziarMF/deep-k-means) (Python + Tensorflow) | [Link](https://clustpy.readthedocs.io/en/latest/clustpy.deep.html#clustpy.deep.dkm.DKM) |
| ENRC | [Deep Embedded Non-Redundant Clustering](https://ojs.aaai.org/index.php/AAAI/article/view/5961) | AAAI 2020 | [Link](https://gitlab.cs.univie.ac.at/lukas/enrcpublic) (Python + PyTorch) | [Link](https://clustpy.readthedocs.io/en/latest/clustpy.deep.html#clustpy.deep.enrc.ENRC) |
| IDEC | [Improved Deep Embedded Clustering with Local Structure Preservation](https://www.ijcai.org/proceedings/2017/243) | IJCAI 2017 | [Link](https://github.com/XifengGuo/IDEC) (Python + Keras) | [Link](https://clustpy.readthedocs.io/en/latest/clustpy.deep.html#clustpy.deep.dec.IDEC) |
| N2D | [N2d:(not too) deep clustering via clustering the local manifold of an autoencoded embedding](https://ieeexplore.ieee.org/document/9413131) | ICPR 2021 | [Link](https://github.com/XifengGuo/IDEC) (Python + Keras) | [Link](https://clustpy.readthedocs.io/en/latest/clustpy.deep.html#clustpy.deep.ddc_n2d.N2D) |
| VaDE | [Variational Deep Embedding: An Unsupervised and Generative Approach to Clustering](https://www.ijcai.org/proceedings/2017/0273) | IJCAI 2017 | [Link](https://github.com/slim1017/VaDE) (Python + Keras) | [Link](https://clustpy.readthedocs.io/en/latest/clustpy.deep.html#clustpy.deep.vade.VaDE) |

# Pyclustering

Clustering algorithms and methods (module pyclustering.cluster):

| **Algorithm** |  |  | **OTHER LINKS** |
| --- | --- | --- | --- |
| Agglomerative | Agglomerative clustering is a type of hierarchical clustering. A foundational reference is: "Methods for Grouping Data" by Joe H. Ward Jr. (1963). Link | Hierarchical clustering is versatile and works well on general-purpose datasets with numerical or categorical attributes. | SCIKIT-LEARN |
| BANG | "The BANG-clustering system: Grid-based data analysis using hierarchical density estimates" by Christian Böhm, Bernhard Braunmüller, et al. (2000). Link | Suitable for datasets with spatial or geometric properties due to its grid-based hierarchical density analysis. |  |
| BIRCH | "BIRCH: An Efficient Data Clustering Method for Very Large Databases" by Tian Zhang, Raghu Ramakrishnan, and Miron Livny (1996). Link | Works on large datasets with numerical attributes, making it a good fit for UCI-like datasets. | SCIKIT-LEARN |
| BSAS | "A model for sequential clustering" by A. G. F. M. Meulman (1982). [Link](https://www.sciencedirect.com/science/article/pii/0377221782900993) | Can handle numerical data effectively in sequential clustering scenarios. |  |
| CLARANS | "CLARANS: A Method for Clustering Objects for Spatial Data Mining" by Raymond T. Ng and Jiawei Han (1994). Link | Optimized for spatial and general clustering problems, often applicable to UCI datasets. |  |
| CLIQUE | "Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications" by Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos, and Prabhakar Raghavan (1998). Link | Suitable for high-dimensional datasets with subspace clustering needs. |  |
| CURE | "CURE: An Efficient Clustering Algorithm for Large Databases" by Sudipto Guha, Rajeev Rastogi, and Kyuseok Shim (1998). Link | Works on large-scale and general-purpose datasets, especially with outlier handling. |  |
| DBSCAN | "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise" by Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu (1996). [Link](https://www.dbs.ifi.lmu.de/Publikationen/Papers/KDD-96.final.frame.pdf) | Well-suited for datasets with arbitrary-shaped clusters and noise. |  |
| Fuzzy C-Means | "Pattern Recognition with Fuzzy Objective Function Algorithms" by J.C. Bezdek (1981). [Link](https://www.springer.com/gp/book/9780306406713) | Works for general-purpose datasets, particularly when soft clustering is needed. |  |
| G-Means | "Learning the k in k-means" by Greg Hamerly and Charles Elkan (2003). Link | Can be applied to classical datasets where the number of clusters is unknown. | **G-Means Clustering Algorithm**  **Article:** "Learning the k in k-means"  **Link:** https://www.cs.utexas.edu/~dml/papers/gmeans.pdf  **Code:GitHub Repository:** [alberto-mate/G-Means](https://github.com/alberto-mate/G-Means) |
| HSyncNet | HSyncNet: A Hybrid Synchronization-based Neural Network for Data Clustering" by Andrei Zinovyev and Fabrice d'Alché-Buc (2001). [Link](https://link.springer.com/chapter/10.1007/3-540-44668-0_27) | Suitable for numerical data but less commonly used in classical datasets. |  |
| K-Means | "Some methods for classification and analysis of multivariate observations" by J. MacQueen (1967). Link | A classic method widely used for numerical datasets. | SCIKIT-LEARN |
| K-Means++ | "k-means++: The Advantages of Careful Seeding" by David Arthur and Sergei Vassilvitskii (2007). Link | An enhancement of K-Means, suitable for numerical datasets. |  |
| K-Medians | "Clustering in metric spaces: A case study using k-median" by Teofilo F. Gonzalez (1985). Link | Effective for numerical data, especially when robustness to outliers is required. |  |
| K-Medoids | "Clustering by means of Medoids" by Leonard Kaufman and Peter J. Rousseeuw (1987). [Link](https://link.springer.com/chapter/10.1007/978-3-642-35903-8_16) | Works well with numerical and categorical data, making it applicable to UCI datasets. |  |
| MBSAS | "Sequential clustering with the Bounded-Self-Organizing Map" by M. Verleysen and D. François (2005). [Link](https://ieeexplore.ieee.org/document/1521425) | Handles numerical data effectively in sequential clustering. |  |
| OPTICS | "OPTICS: Ordering Points To Identify the Clustering Structure" by Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, and Jörg Sander (1999). Link | Useful for discovering clusters in datasets with varying density. |  |
| ROCK | "ROCK: A Robust Clustering Algorithm for Categorical Attributes" by Sudipto Guha, Rajeev Rastogi, and Kyuseok Shim (1999). Link | Optimized for categorical data and can be applied to certain UCI datasets. |  |
| SOM-SC | "SOM-SC: Self-Organizing Map for Subspace Clustering" by Markus M. Breunig, Hans-Peter Kriegel, and Jörg Sander (2000). [Link](https://link.springer.com/chapter/10.1007/3-540-44471-6_27) | Suitable for numerical data and subspace clustering in classical datasets. |  |
| SyncNet | "SyncNet: A Synchronization-based Neural Network for Clustering" by Andrei Zinovyev and Fabrice d'Alché-Buc (2000). [Link](<https://link.springer.com/chapter/10.1007/3-> | Works on numerical data but is less common for standard UCI datasets. |  |
| Sync-SOM | [A Synchronization-based Self-Organizing Map (d'Alché-Buc & Zinovyev, 2001)](https://link.springer.com/chapter/10.1007/3-540-44668-0_27) | Similar to SyncNet, applicable for numerical data clustering. |  |
| TTSAS |  |  |  |
| X-Means | X-Means: Extending K-Means with Efficient Estimation of the Number of Clusters (Pelleg & Moore, 2000) | Extends K-Means for unknown cluster counts and is effective on numerical data. | **Article:** "Extending K-means with Efficient Estimation of the Number of Clusters"  **Link:**<https://www.cs.cmu.edu/~dpelleg/download/xmeans.pdf>  **Code: GitHub Repository:** [KazuhisaFujita/X-means](https://github.com/KazuhisaFujita/X-means) |

Affinity

<https://scikit-learn.org/1.5/modules/clustering.html#affinity-propagation>

MeanShift

<https://scikit-learn.org/1.5/modules/clustering.html#mean-shift>

SpectralClustering

<https://scikit-learn.org/1.5/modules/clustering.html#spectral-clustering>

**Article:** "Web-Scale K-Means Clustering"

**Link:** <https://www.eecs.tufts.edu/~dsculley/papers/fastkmeans.pdf>

**Code:** [**https://scikit-learn.org/dev/modules/generated/sklearn.cluster.MiniBatchKMeans.html**](https://scikit-learn.org/dev/modules/generated/sklearn.cluster.MiniBatchKMeans.html)