IEEE WCCI 2020

IEEE World Congress on Computational Intelligence

Virtual Conference – July 19-24, 2020

FDBSCAN-APT
A Fuzzy Density-based Clustering Algorithm with Automatic Parameter Tuning

Alessio **Bechini** Martina **Criscione** Pietro **Ducange**
Francesco **Marcelloni** Alessandro **Renda**

University of Pisa, Dept. of Information Engineering

Outline

- Intro: DBSCAN
- FDBSCAN-APT: Motivation and Goals
- A fuzzy extension of DBSCAN clustering algorithm
- A novel heuristic for Automatic Parameter Tuning
- Experimental Setup and Results
- Conclusions

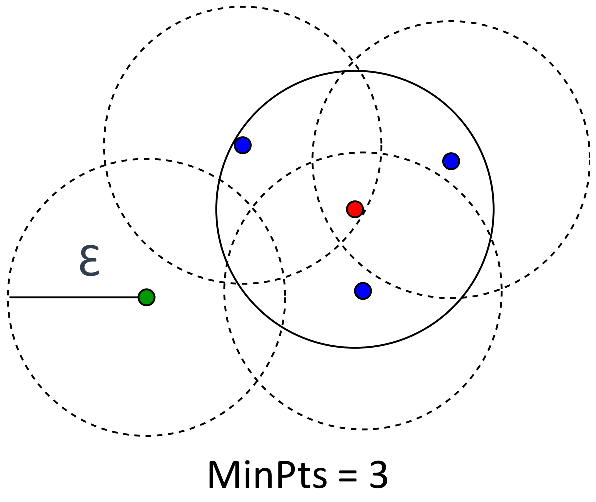
Intro: DBSCAN

Partitions data into **connected dense regions** separated by sparse regions

- Distinction between Core, Border, Noise objects

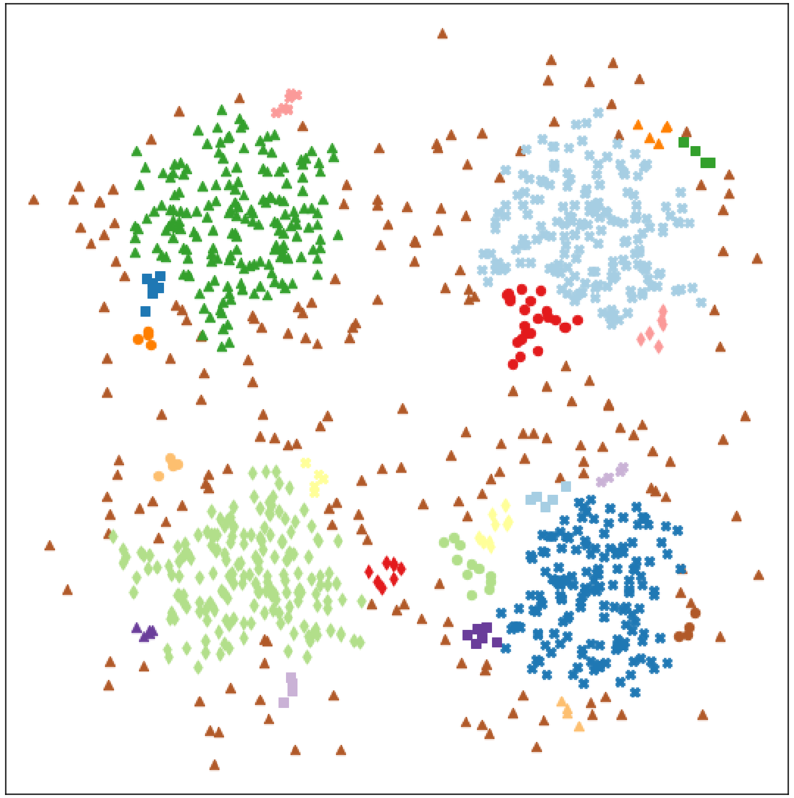
Requires the definition of **two input parameters**

- ε: defines the neighborhood size
- **MinPts**: minimum number of objects required for a core

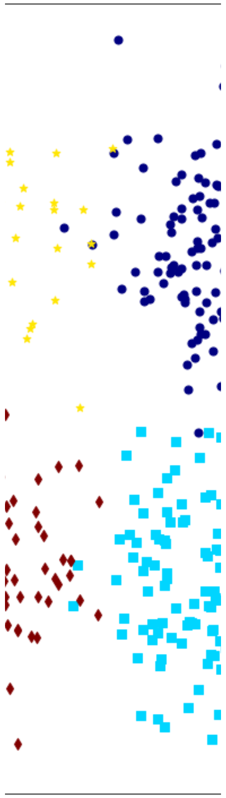


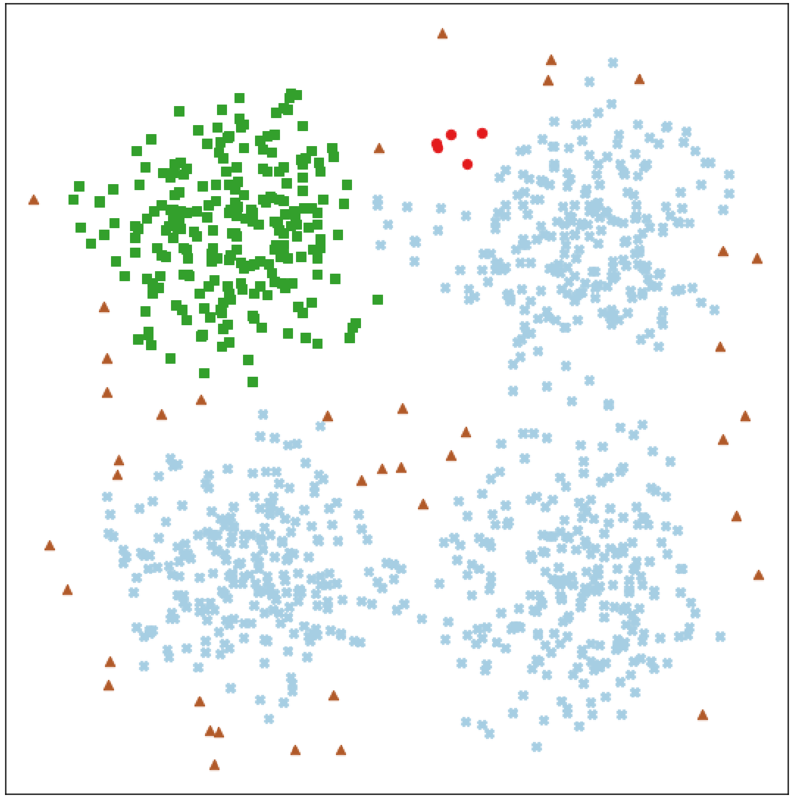
- Can discover clusters with arbitrary shapes
- Does not require prior knowledge of the number of clusters
- Crucial importance of **input parameter setting**

A (simple?) Clustering Task



MinPts, small **ε**



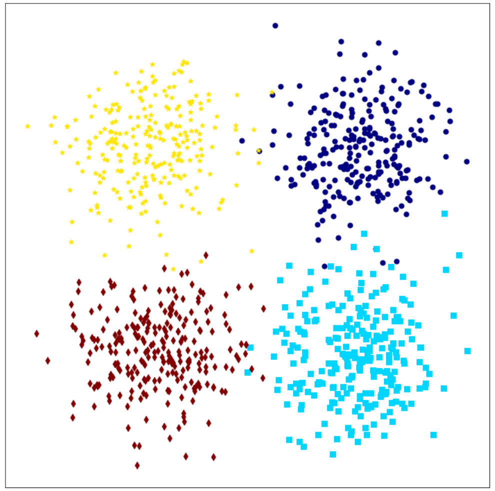


MinPts, big **ε**

FDBSCAN-APT: Goals

On this example, we can draw general goals

- Discover clusters with **fuzzy overlapping borders**
- Automatically find **proper values of input parameters**

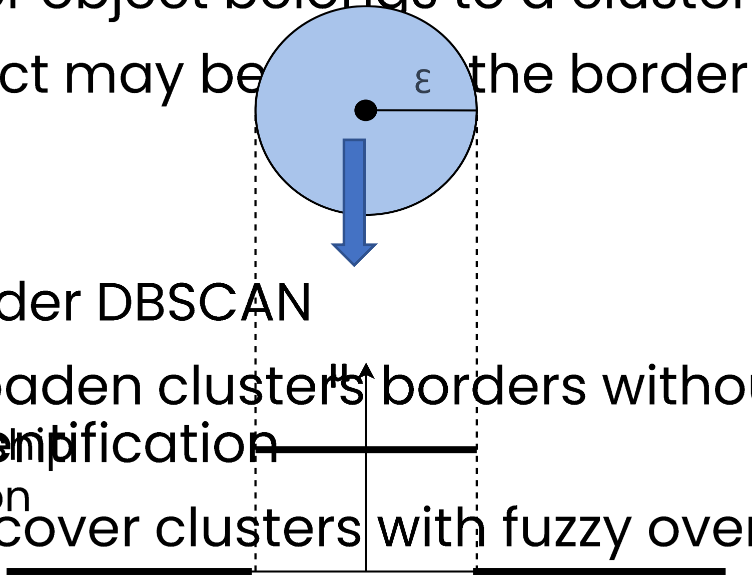


FDBSCAN APT
Fuzzy DBSCAN with Automatic Parameter Tuning

Fuzzy Border DBSCAN

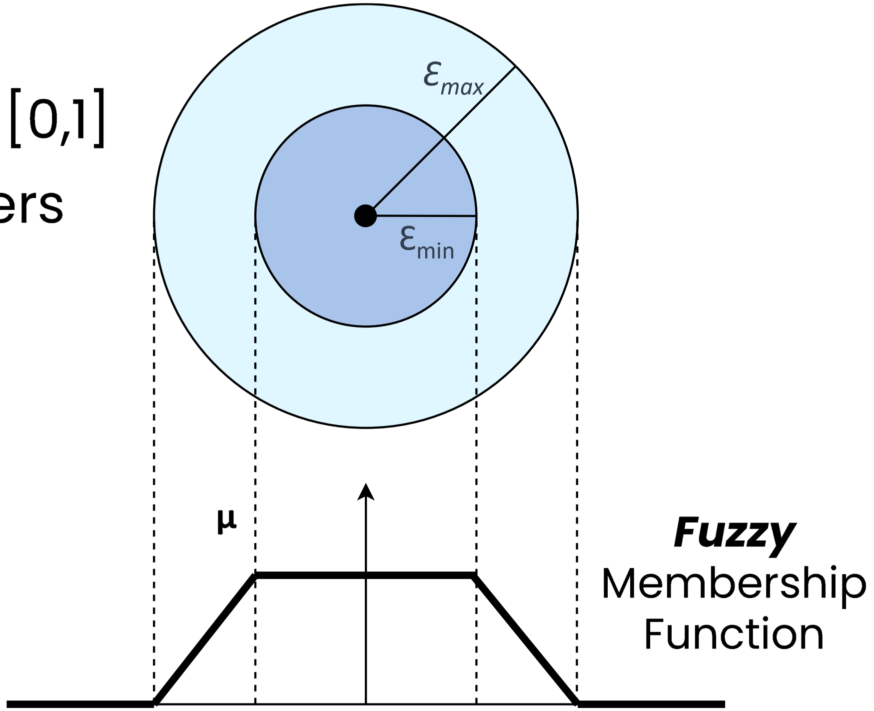
Basic idea: **fuzzy membership function**

- Only objects within ε_{min} are considered for the determination of a ***core object***
- A border object belongs to a cluster with a degree in [0,1]
- An object may be … the border of multiple clusters



Fuzzy border DBSCAN

- can broaden clusters' borders without affecting core identification
- can discover clusters with fuzzy overlapping borders



Ienco D. and Bordogna G. "Fuzzy extensions of the DBScan clustering algorithm." Soft Computing (2018)

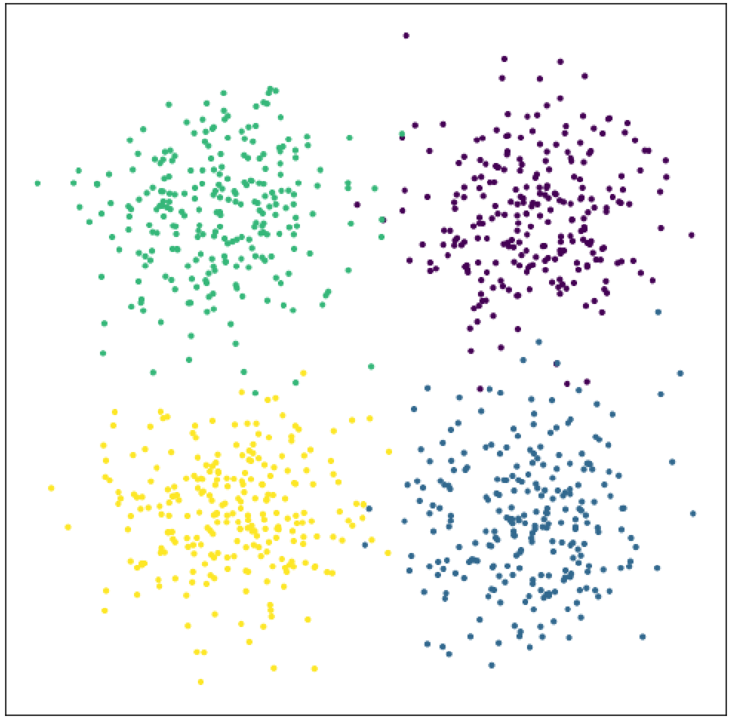
Automatic Parameter Tuning

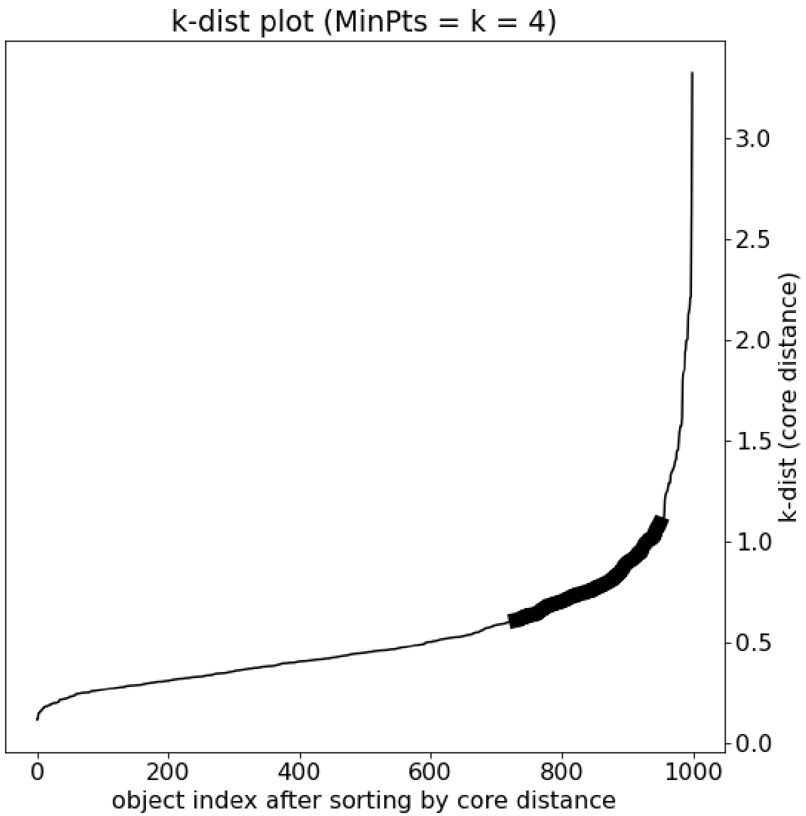
Proposed approach

- Fix $MinPts$
- Estimate ε_{min} and ε_{max}

Basic idea

- Resort to the **k-dist plot**
  - Evaluate the *core-distance* for each object
  - Plot after sorting

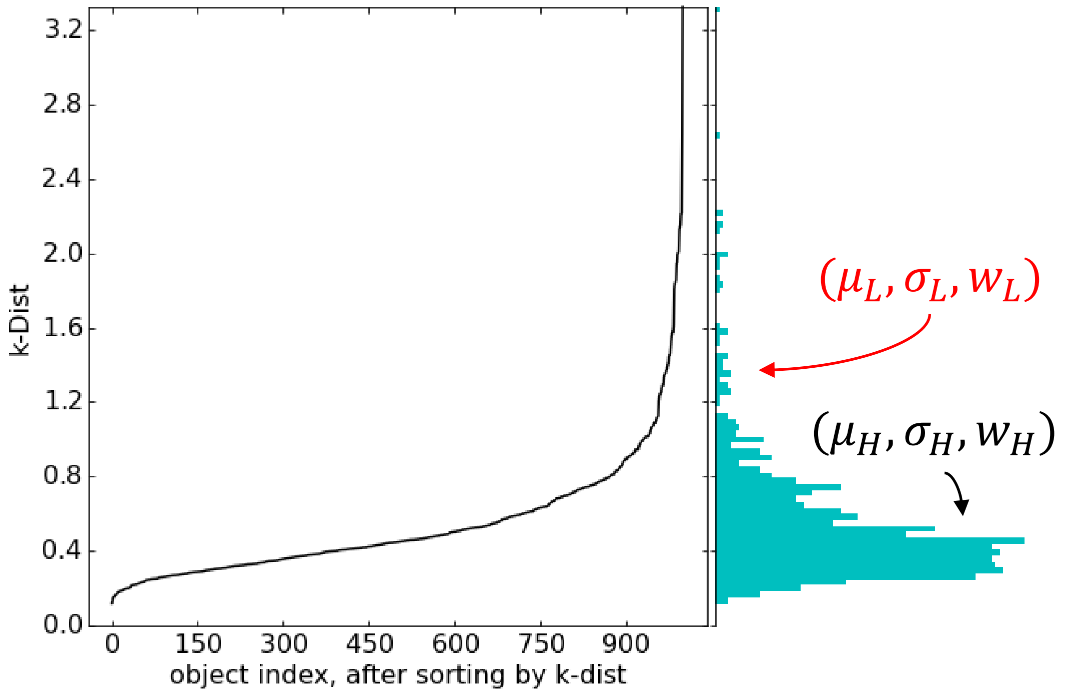




Automatic Parameter Tuning

Two assumptions:

- The dataset distribution is **unimodal**
    - all clusters have roughly the same density of objects
- The array of core-distances can be modeled as **a mixture of two Gaussian components**
    - the first one models the contribution of objects within a high-density region
    - the second one models the contribution of border objects and is affected by the presence of noise and outliers



Automatic Parameter Tuning

- Given the Gaussian Mixture Model fitting on the **k-dist** array
  - (μ_H, σ_H) the parameters of the *high-density* Gaussian component
  - (μ_L, σ_L) the parameters of the *low-density* Gaussian component

- A heuristic for **Automatic FDBSCAN parameter setting**
  - $MinPts = k$
  - $\hat{\varepsilon}_{min} = \mu_H + 2\,\sigma_H$
  - $\hat{\varepsilon}_{max} = \alpha * \hat{\varepsilon}_{min} * \frac{\mu_L}{\mu_L - \mu_H}$

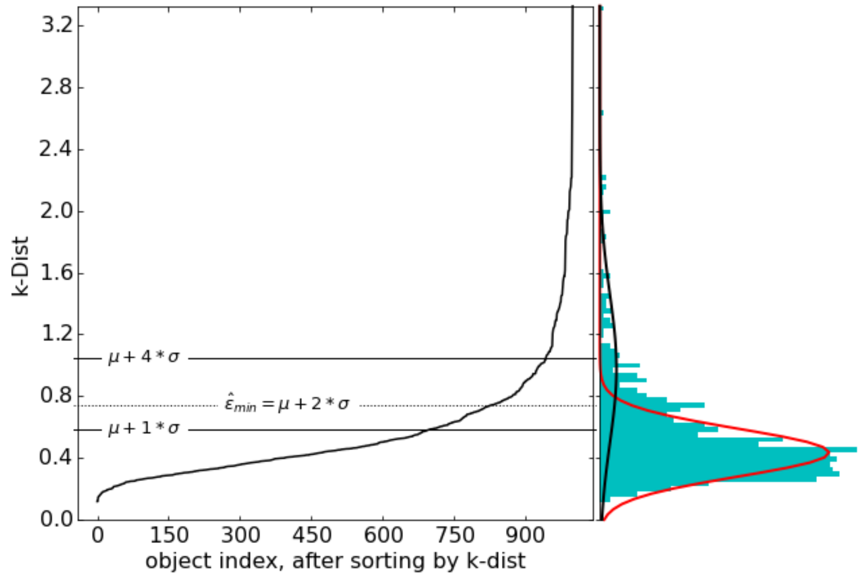
- Approximately 98% of objects of the high-density component will meet the core condition
- Expressed as a function of $\hat{\varepsilon}_{min}$
- User defined parameter $\alpha \geq 1$ for flexibility
- Last coefficient for *narrowing borders* in presence of noise

Experimental Setup

Comparison of **FDBSCAN-APT**
with **50** other **parameter configurations** of FDBSCAN

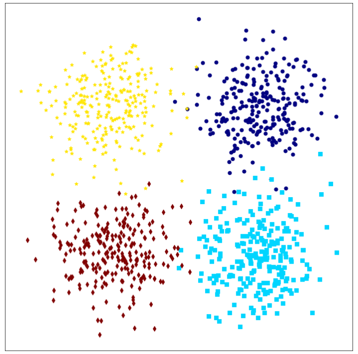
ε_{min}: 10 evenly spaced values in the range $[\mu_H + \sigma_H,\ \mu_H + 4\ \sigma_H]$

ε_{max}: 5 evenly spaced values in the range $[\varepsilon_{min},\ 5 * \varepsilon_{min}]$

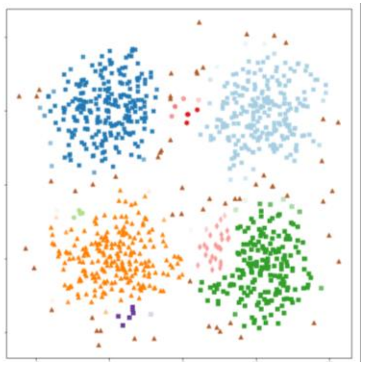


- Nine bidimensional synthetic datasets
- Clustering results evaluation in terms of **Adjusted Rand Index**

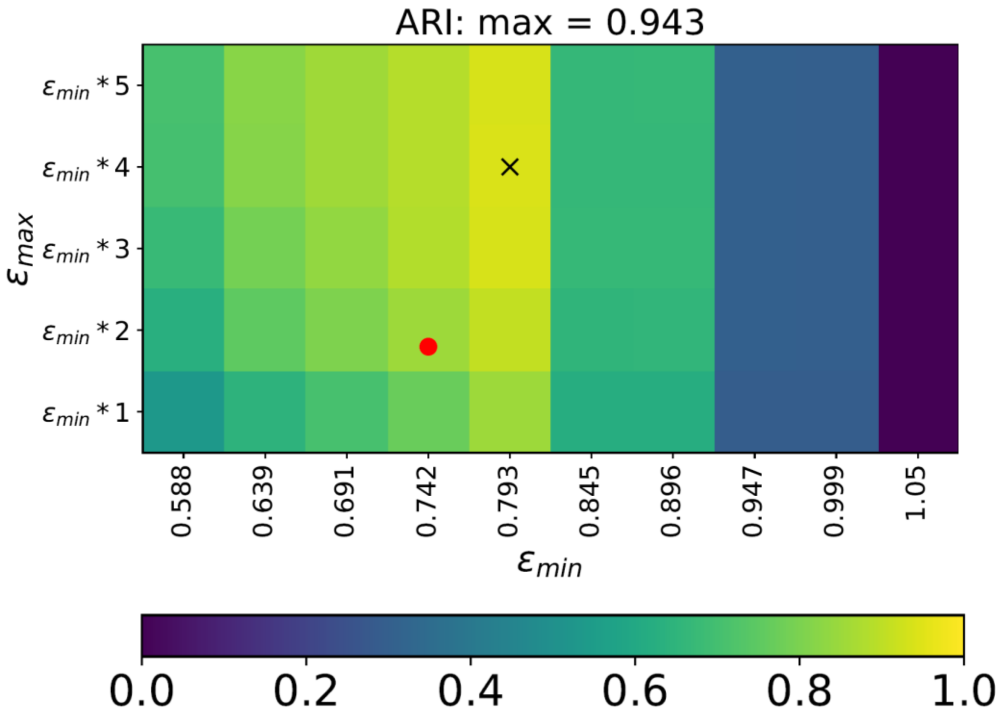
Experimental Results



Square1



FDBSCAN-APT output

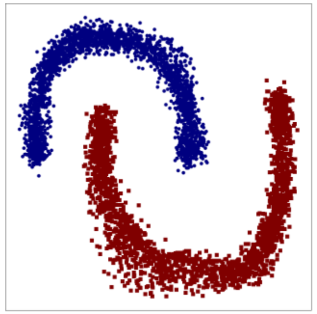


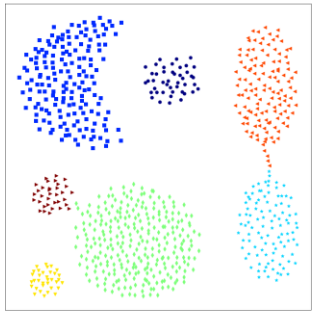
● FDBSCAN-APT

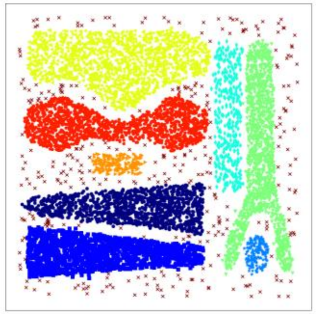
x Grid best configuration

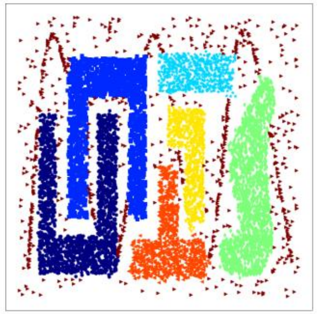
- Parameter setting is **crucial**
- FDBSCAN-APT automatically finds an ***acceptable* configuration**
- **Introduction of fuzziness** is beneficial for modeling this dataset

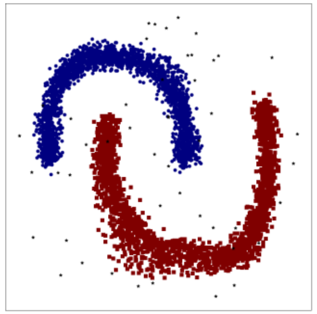
Experimental Results

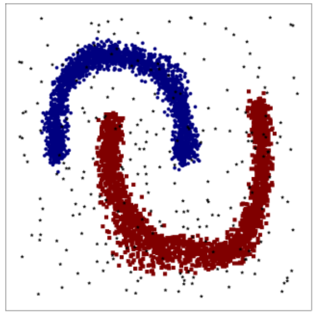

Banana

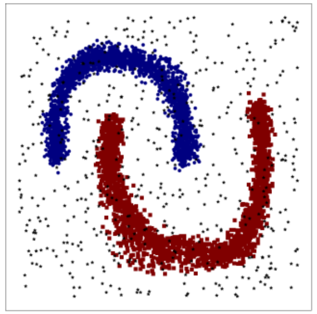

Aggregation

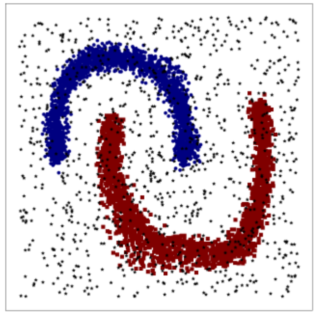

Cluto-t8-8k


Cluto-t4-8k


Banana_noise_1

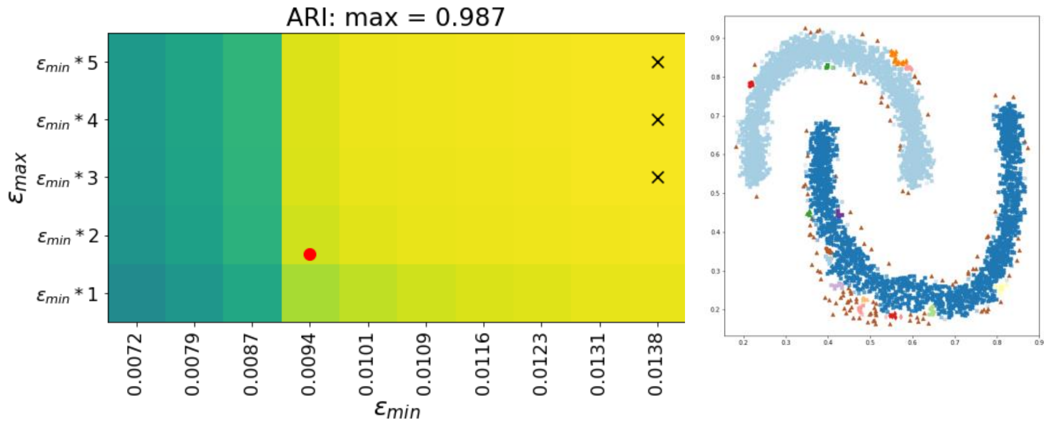

Banana_noise_5

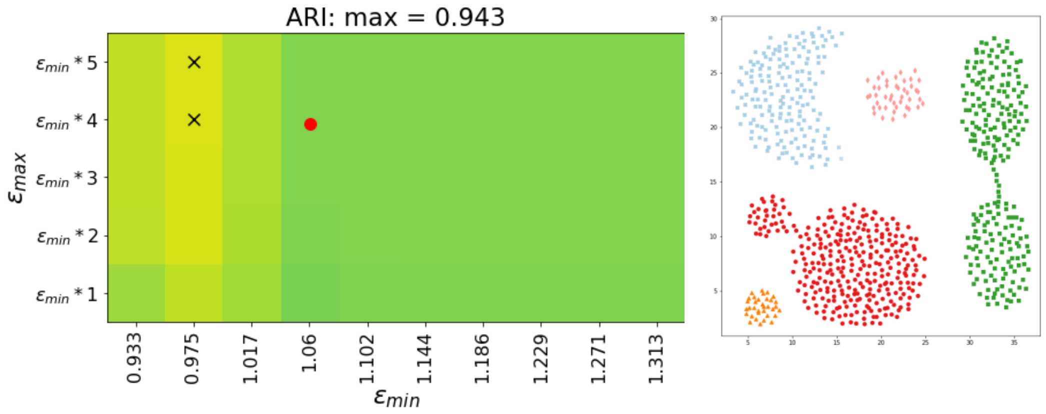

Banana_noise_10

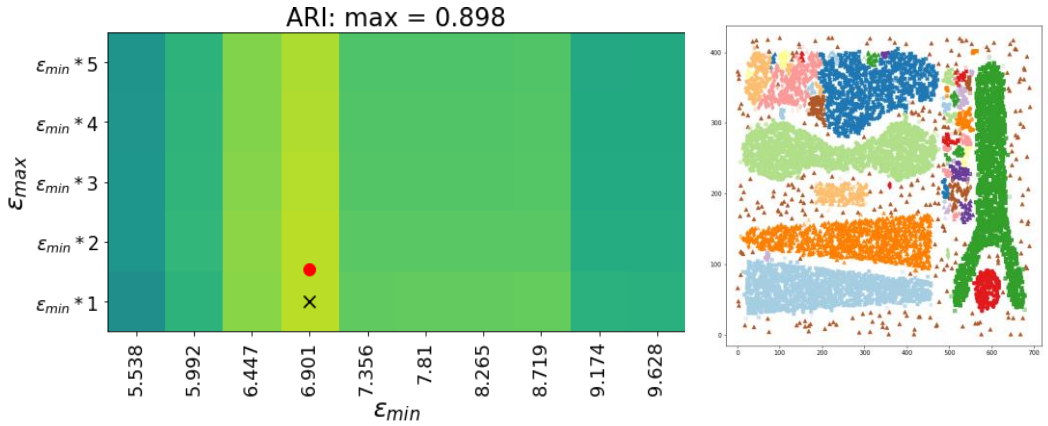

Banana_noise_20

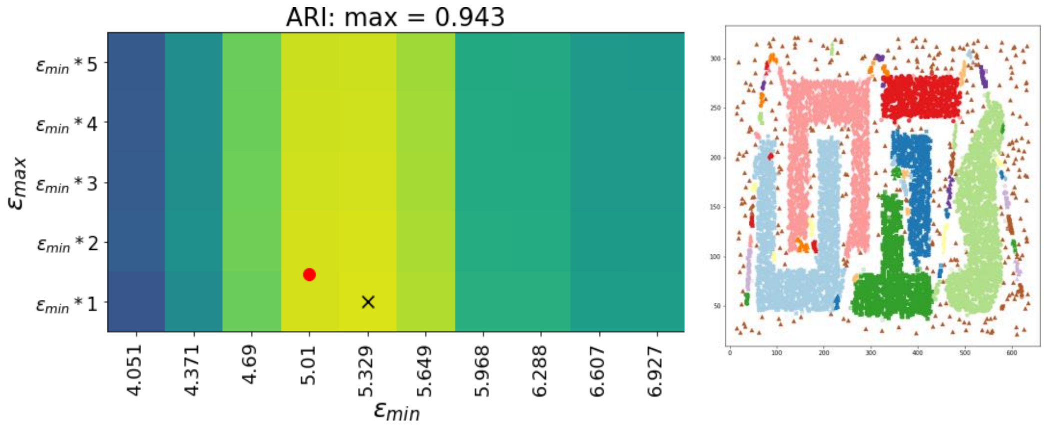
Clustering Results: ARI		
Dataset	Grid best	FDBSCAN-APT
Square1	0.943	0.844
Banana	0.986	0.916
Banana_noise_1	0.989	0.931
Banana_noise_5	0.956	0.908
Banana_noise_10	0.930	0.882
Banana_noise_20	0.866	0.840
Cluto-t4-8k	0.943	0.943
Cluto-t8-8k	0.898	0.903
Aggregation	0.943	0.809

Experimental Results

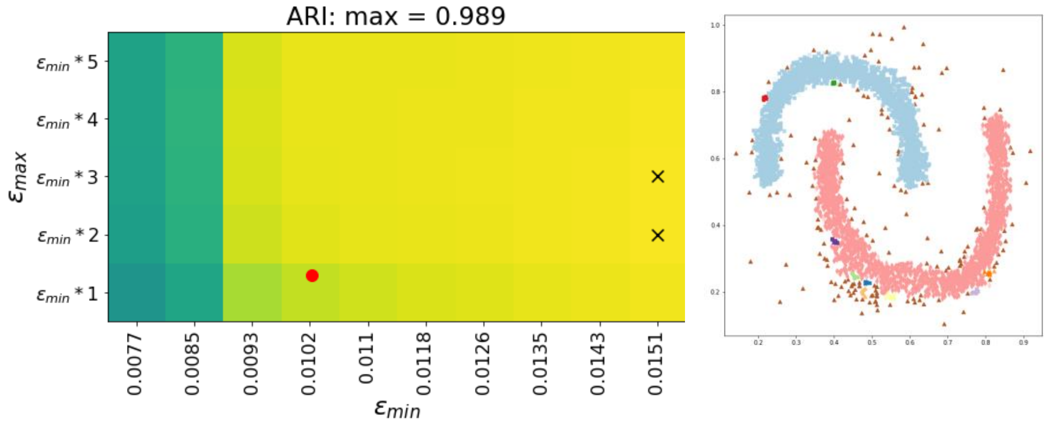


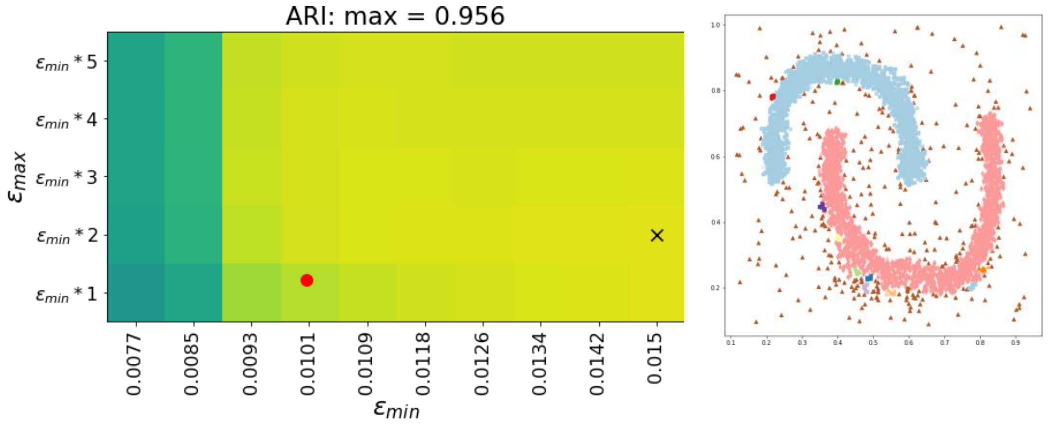


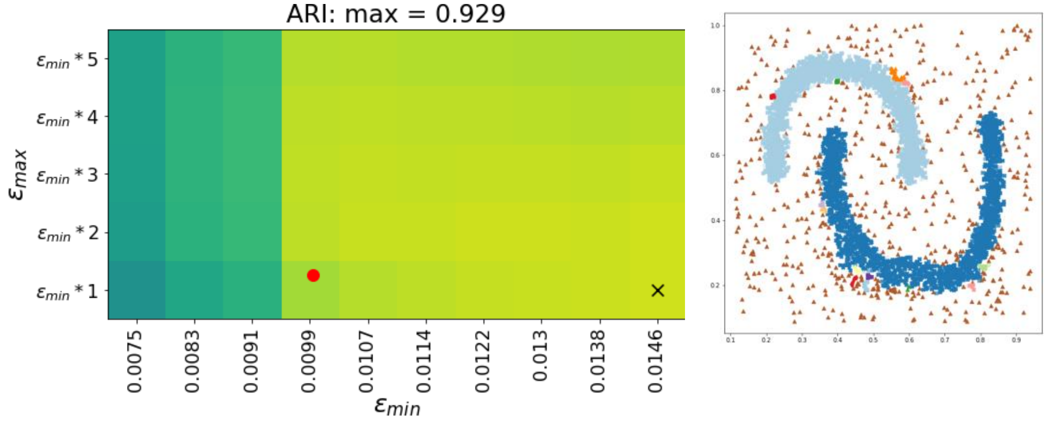


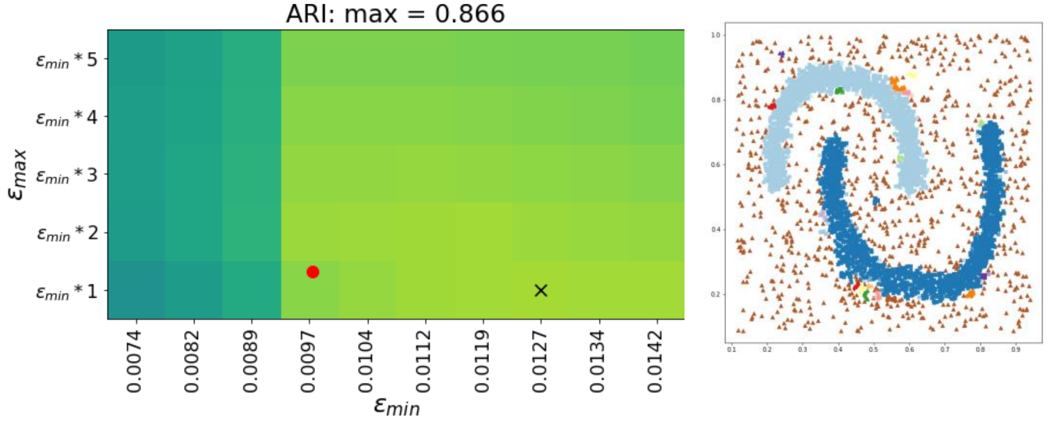


Experimental Results









Conclusions

Proposal of **FDBSCAN-APT** clustering algorithm

- It enables the detection of clusters with **fuzzy overlapping borders**
- A novel heuristic proposed for **Automatic Parameter Tuning** addresses the crucial issue of input parameters setting
- Effectiveness of the proposed approach is shown on several **synthetic datasets**

Towards further developments

- Evaluation on real/big/high-dimensional datasets
- Extension to multi-density datasets

Thank you for your attention

Alessandro Renda
Smart Computing Ph.D. student
alessandro.renda@unifi.it
University of Pisa, Dept. Information Engineering
Computational Intelligence Group