DCF: An Efficient and Robust Density-Based Clustering Method

Joshua Tobin & Mimi Zhang

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Problem Statement

Introduction

Mode-seeking clustering associates each point to a mode of the underlying probability density function.

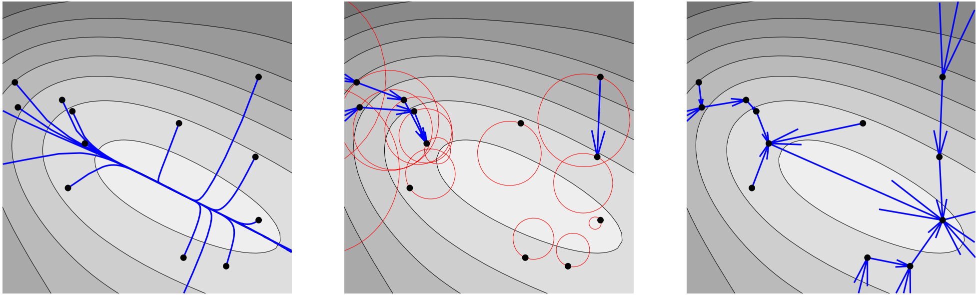


Figure: 1 of Vedaldi and Soatto 2008.

Problem Statement

Introduction

Mode-seeking clustering associates each point to a mode of the underlying probability density function.

Benefits:

- Detect clusters with arbitrary structure.
- Number of clusters not required as an input.

Challenges:

- Point modes are not robust.
- Parameter tuning is hard to assess.
- High computational complexity.

Can we develop a fast, flexible and robust mode-seeking method?

Our Contribution

Introduction

We introduce Density Core Finding (DCF) aiming at improving the applicability and efficiency of Density Peaks Clustering (DPC).

By directing the peak-finding method to detect modal sets, our algorithm is:

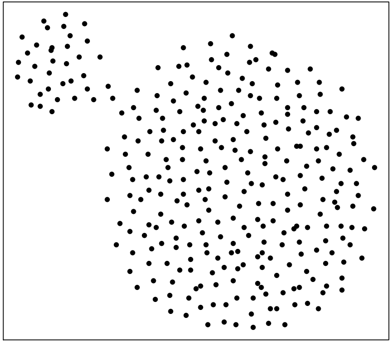
1. applicable to large datasets,
2. capable of detecting clusters of varying density,
3. competent at deciding the correct number of clusters.

Density Peaks Clustering

Preliminaries

> "Cluster centers ..are surrounded by neighbors with lower local density ..and they are at a relatively large distance from any points with a higher local density."
>
> \- *Rodriguez and Laio 2014*

- DPC is a mode-seeking clustering algorithm.
- Cluster centers are selected using a heuristic known as the peak-finding criterion.
- Non-center instances are assigned to clusters using a hill-climbing method.



Density Peaks Clustering

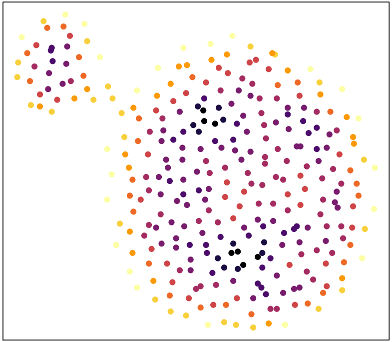
Preliminaries

> "Cluster centers ..are surrounded by neighbors with lower local density ..and they are at a relatively large distance from any points with a higher local density."
>
> \- *Rodriguez and Laio 2014*

For every $x \in \mathbb{R}^d$, let d_j be the distance from x to x_j.

The density estimate is given as

$$f(x) := \sum_j \mathbb{1}\left(\text{distance } d_j < \text{distance } d_c\right).$$



Density Peaks Clustering

Preliminaries

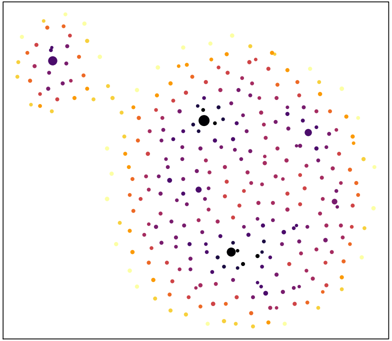
"Cluster centers ..are surrounded by neighbors with lower local density ..and they are at a relatively large distance from any points with a higher local density."

- *Rodriguez and Laio 2014*

$\delta(x)$ is the distance to the nearest neighbor of higher local density.

We define the peak-finding criterion as

$$\gamma(x) = f(x) \cdot \delta(x).$$



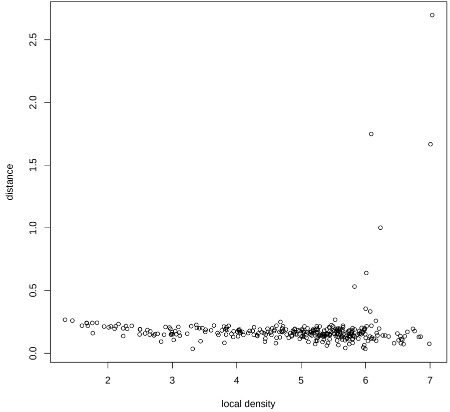
Density Peaks Clustering

Preliminaries

> "Cluster centers ..are surrounded by neighbors with lower local density ..and they are at a relatively large distance from any points with a higher local density."
>
> - *Rodriguez and Laio 2014*

In the work of Rodriguez & Laio, cluster centers are selected manually from the decision graph:

$$\{(f(x), \delta(x)) : x \in \boldsymbol{X}\}.$$



Density Peaks Clustering

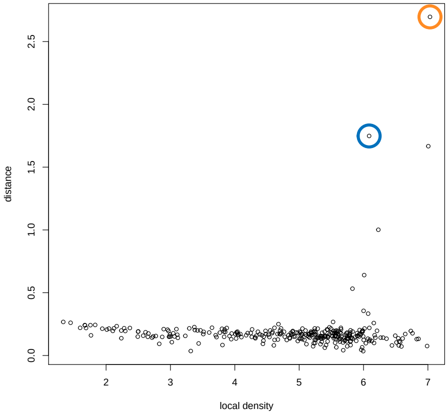
Preliminaries

"Cluster centers ..are surrounded by neighbors with lower local density ..and they are at a relatively large distance from any points with a higher local density."

- *Rodriguez and Laio 2014*

In the work of Rodriguez & Laio, cluster centers are selected manually from the decision graph:

$$\{(f(x), \delta(x)) : x \in \boldsymbol{X}\}.$$

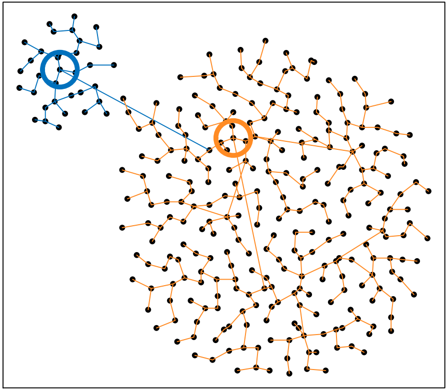


Density Peaks Clustering

Preliminaries

> "Cluster centers ..are surrounded by neighbors with lower local density ..and they are at a relatively large distance from any points with a higher local density."
>
> \- *Rodriguez and Laio 2014*

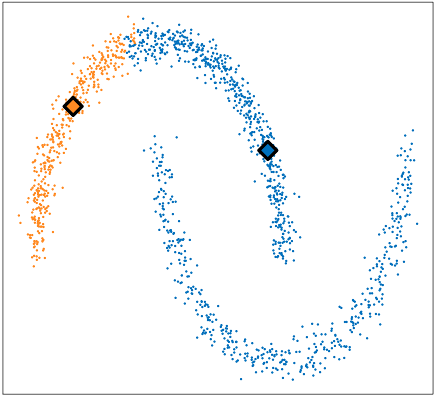
Finally, all non-center instances are assigned to the same cluster as their nearest neighbor of higher density.



Cluster Cores

Preliminaries

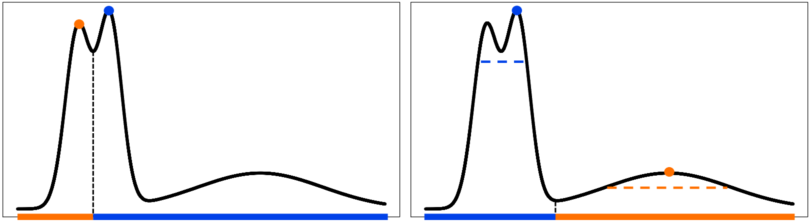
Problem: DPC regularly fails when data contains both high- and low-density clusters.



1. Peak-finding criterion erroneously selects multiple centers from high-density clusters.
2. The allocation mechanism incorrectly assigns all points in the low-density cluster.

Cluster Cores

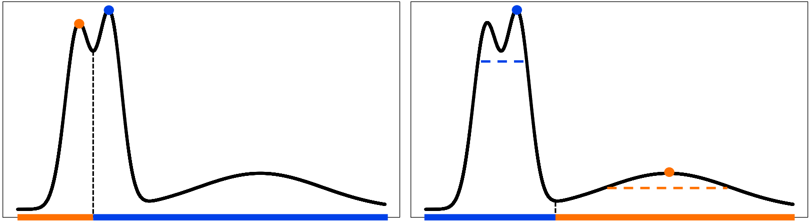
Preliminaries



- Jiang and Kpotufe 2017; Jiang, Jang, and Kpotufe 2018 develop the notion of modal-sets for methods MCores and QuickShift++.
- Model locally high-density regions of the data with sets of arbitrary shape, size and density level.
- Parametrized by $\beta \in (0, 1)$, determining how much the density can fluctuate within a cluster.

Cluster Cores

Preliminaries



- For each instance x^* with local density $f(x^*) = \lambda^*$, an associated level set is found

$$\chi = \{x \in \boldsymbol{X} : f(x) \geq \lambda^* - \beta\lambda^*\}.$$

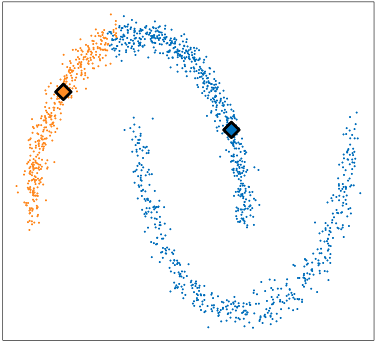
- If the subset of χ containing x^* is disconnected from all previous modal-sets, it is accepted as a cluster core.

Density Core Finding

Our Proposal

Solution: Direct the peak-finding criterion to detect modal-sets.

- Reduces risk of selecting multiple centers from high-density cluster.
- Less sensitive to chance variation in empirical density estimate.

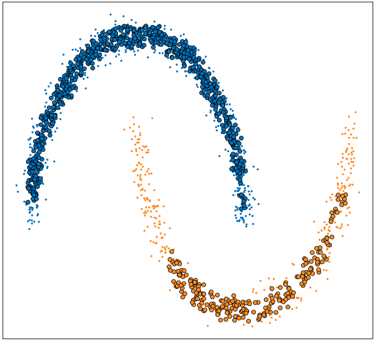


Density Core Finding

Our Proposal

Solution: Direct the peak-finding criterion to detect modal-sets.

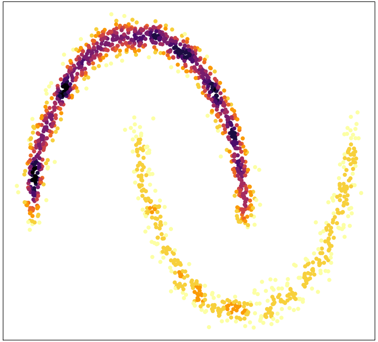
- Reduces risk of selecting multiple centers from high-density cluster.
- Less sensitive to chance variation in empirical density estimate.



The Algorithm

Our Proposal

DCF Algorithm: Detects clusters of arbitrary shape, size and density automatically and executes in $O(n \log n)$.



For every $x \in \mathbb{R}^d$, let $r_k(x)$ denote the distance from x to its k-th nearest neighbor.

The density estimate is given as

$$f_k(x) := \frac{k}{n \cdot Vol(B(x, r_k(x)))},$$

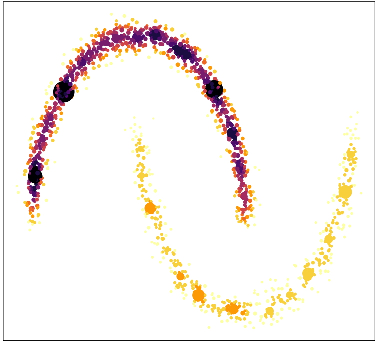
where v_d is the volume of the unit sphere in $\mathbb{R}^d$.

Joshua Tobin & Mimi Zhang

The Algorithm

Our Proposal

DCF Algorithm: Detects clusters of arbitrary shape, size and density automatically and executes in $O(n \log n)$.



$\delta_k(x)$ is the distance to the nearest neighbor of higher local density.

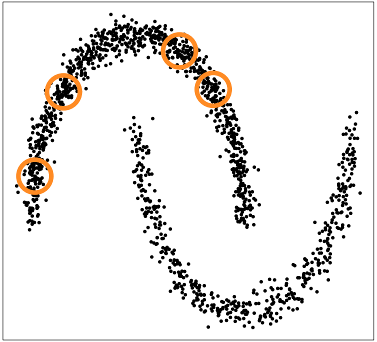
The peak-finding criterion is

$$\gamma_k(x) = f_k(x) \cdot \delta_k(x).$$

The Algorithm

Our Proposal

DCF Algorithm: Detects clusters of arbitrary shape, size and density automatically and executes in $O(n \log n)$.



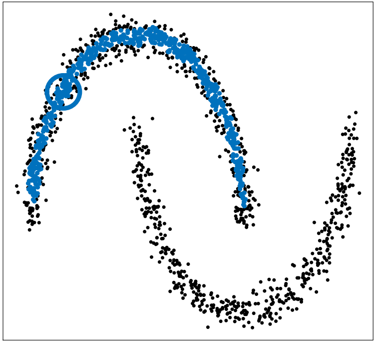
Selecting instances with maximal value of γ_k gives incorrect clustering.

Four points from high-density cluster have larger γ_k than the max in low-density cluster.

The Algorithm

Our Proposal

DCF Algorithm: Detects clusters of arbitrary shape, size and density automatically and executes in $O(n \log n)$.



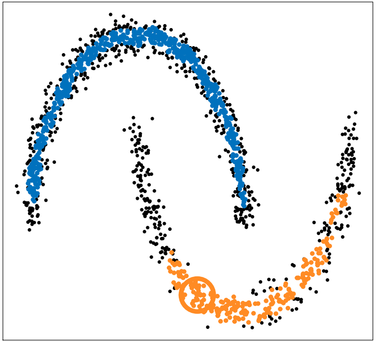
DCF finds $x = \arg\max_{x \in \boldsymbol{X}} \gamma_k(x)$ and sets $\lambda := f_k(x)$.

Taking $A_\beta(x)$ to be the set of points connected to x, we add $A_\beta(x)$ to the set of cluster cores and mark all points in $A_\beta(x)$ as assessed.

The Algorithm

Our Proposal

DCF Algorithm: Detects clusters of arbitrary shape, size and density automatically and executes in $O(n \log n)$.



Find $x = \arg\max_{x \in \mathbf{X}} \{\gamma_k(x) : x \notin \mathit{Assessed}\}$.

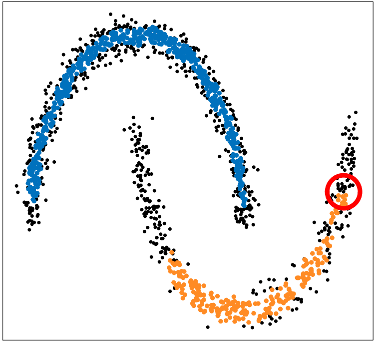
Set $\lambda := f_k(x)$, identify the points connected to x as $A_\beta(x)$and mark all points in $A_\beta(x)$ as assessed.

If $A_\beta(x)$ is disjoint from all cluster cores, add $A_\beta(x)$ to $\widehat{\mathcal{M}}$.

The Algorithm

Our Proposal

DCF Algorithm: Detects clusters of arbitrary shape, size and density automatically and executes in $O(n \log n)$.

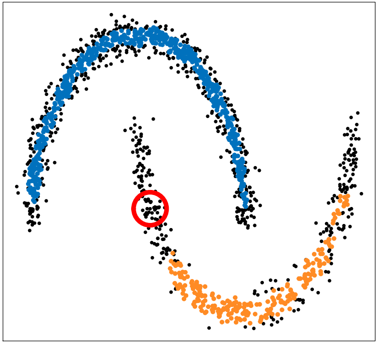


The search procedure terminates when every $x \in \boldsymbol{X}$ has been assessed.

The Algorithm

Our Proposal

DCF Algorithm: Detects clusters of arbitrary shape, size and density automatically and executes in $O(n \log n)$.

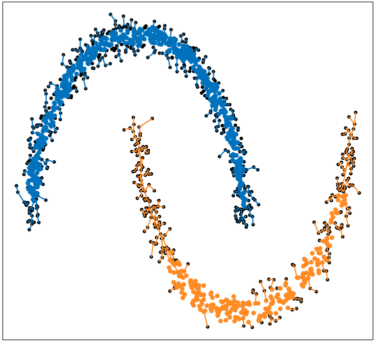


The search procedure terminates when every $x \in \boldsymbol{X}$ has been assessed.

The Algorithm

Our Proposal

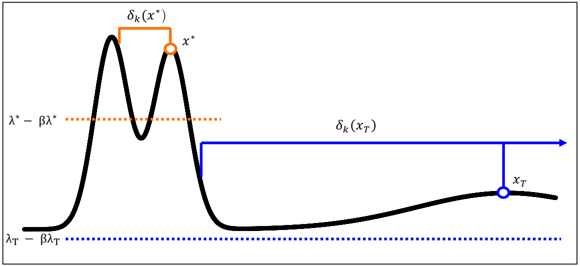
DCF Algorithm: Detects clusters of arbitrary shape, size and density automatically and executes in $O(n \log n)$.



Finally, all non-core instances are assigned to the same cluster as their nearest neighbor of higher density.

Mode Recovery

Analysis



The point x^* is assessed iff. $\gamma_k(x^*) > \gamma_k(x_T)$.

As $\delta_k(x_T)$ is not bounded, we cannot guarantee all modes will be found.

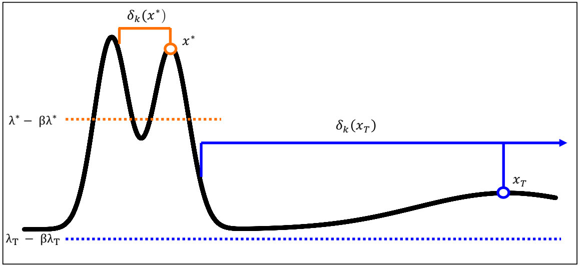
Theoretical results demonstrate why this is unlikely to hinder performance.

Proposition 1

Any cluster that corresponds to a connected component in the mutual k-NN graph will be recovered by DCF.

Mode Recovery

Analysis



The point x^* is assessed iff. $\gamma_k(x^*) > \gamma_k(x_T)$.

As $\delta_k(x_T)$ is not bounded, we cannot guarantee all modes will be found.

Theoretical results demonstrate why this is unlikely to hinder performance.

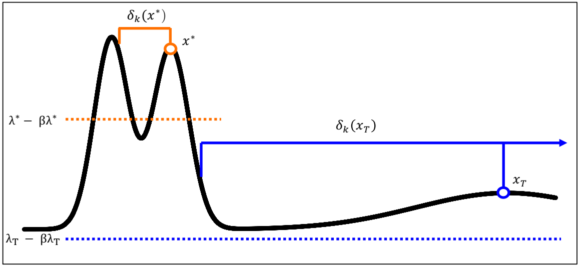
Proposition 2

The probability x_T being connected to the remainder of the graph decreases as the magnitude of $\delta_k(x_T)$ increases.

[Adapted from Prop. 6 of Maier, Hein, and von Luxburg 2009]

Mode Recovery

Analysis



The point x^* is assessed iff. $\gamma_k(x^*) > \gamma_k(x_T)$.

As $\delta_k(x_T)$ is not bounded, we cannot guarantee all modes will be found.

Theoretical results demonstrate why this is unlikely to hinder performance.

Proposition 3

If DCF terminates at x_T with termination density level $\lambda_T - \beta\lambda_T$, $\lambda_T - \beta\lambda_T$ is at least as low as the lowest dip in density between clusters in $\boldsymbol{X}$.

Set Up

Experiments

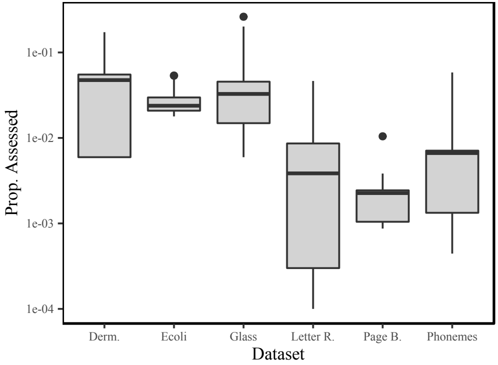
- We compare the performance of DCF with:
  - QuickShift++ (QSP)
  - Density Peaks Clustering (DPC)
  - Adaptive DPC (ADP)
  - Comparative DPC (CDP)
  - DBSCAN (DBS)
  - HDBSCAN (HDB)
- DCF is assessed on six real-world datasets, five UCI datasets and the Phonemes dataset.
- The clusterings are assessed using Adjusted Rand Index (ARI) and Adjusted Mutual Information (AMI) as well as the time taken to execute.

Project Github Repository:

- Implementation of DCF in Python
- Code to replicate all experiments
- `https://github.com/tobinjo96/DCFcluster`

Dataset	Metric	DCF	QSP	DPC	ADP	CDP	DBS	HDB
Derm.	ARI	0.72	0.70	0.22	0.59	**0.73**	0.44	0.47
	AMI	**0.78**	**0.78**	0.45	0.73	0.75	0.63	0.66
Ecoli	ARI	**0.73**	**0.73**	0.55	0.72	0.51	0.50	0.40
	AMI	**0.68**	**0.68**	0.50	0.65	0.55	0.48	0.41
Glass	ARI	**0.31**	0.30	0.20	0.26	0.25	0.25	0.25
	AMI	**0.42**	0.40	0.27	0.38	0.38	0.38	0.37
Letter R.	ARI	0.20	0.20	**0.22**	0.10	0.13	0.07	0.02
	AMI	**0.59**	0.58	0.53	0.33	0.42	0.46	0.45
Page B.	ARI	**0.46**	**0.46**	0.39	0.38	0.42	0.32	0.33
	AMI	**0.30**	**0.30**	0.27	0.26	0.29	0.18	0.20
Phonemes	ARI	**0.76**	**0.76**	0.71	0.70	0.56	0.44	0.36
	AMI	**0.83**	0.80	0.79	0.75	0.66	0.61	0.57

Dataset	DCF	QSP	DPC	ADP	CDP	DBS	HDB
Derm.	0.07	0.03	4.65	2.5	0.32	**0.01**	0.02
Ecoli	0.06	0.02	2.54	1.67	0.13	**0.00**	0.02
Glass	0.03	0.05	0.61	0.24	0.11	**0.00**	**0.00**
Letter R.	**12.79**	19.21	2430.84	1002.42	372.14	19.94	25.53
Page B.	0.73	1.61	123.27	43.26	14.59	1.23	**0.68**
Phonemes	**7.21**	8.79	1627.81	57.33	43.22	15.26	11.42



Face Detection

Application

> *"Modern clustering problems require efficient detection of clusters with arbitrary shape, size and density."*

- Face recognition is a central problem in computer vision.
- We apply DCF to numerical features extracted from two prominent face datasets.

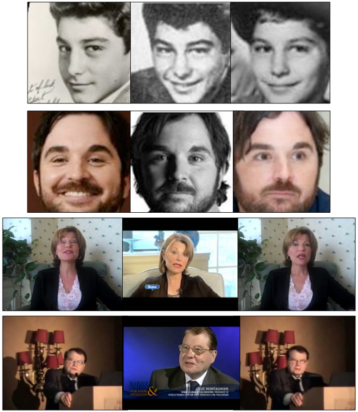
Name	**Instances**	**Dim**	**Identities**
MS-Celeb-1M	1,160,507	256	17,146
YTB-Faces	155,282	256	1,595

Face Detection

Application

Dataset	Metric	DCF	QSP	OPT
MS-Celeb	ARI	**0.90**	0.83	-
	AMI	**0.96**	0.92	-
YTB-Faces	ARI	**0.69**	0.52	0.06
	AMI	**0.91**	0.88	0.15

Dataset	DCF	QSP	OPT
MS-Celeb	**13202.00**	39212.14	-
YTB-Faces	**2212.59**	4338.95	29631.24



Bibliography I

Jiang, Heinrich, Jennifer Jang, and Samory Kpotufe (July 2018). "Quickshift++: Provably Good Initializations for Sample-Based Mean Shift". In: *International Conference on Machine Learning*. PMLR, pp. 2294–2303.

Jiang, Heinrich and Samory Kpotufe (Apr. 2017). "Modal-Set Estimation with an Application to Clustering". In: *Artificial Intelligence and Statistics*. PMLR, pp. 1197–1206.

Maier, Markus, Matthias Hein, and Ulrike von Luxburg (Apr. 2009). "Optimal Construction of K-Nearest-Neighbor Graphs for Identifying Noisy Clusters". In: *Theoretical Computer Science*. Algorithmic Learning Theory 410.19, pp. 1749–1764. ISSN: 0304-3975. DOI: 10.1016/j.tcs.2009.01.009.

Rodriguez, Alex and Alessandro Laio (2014). "Clustering by Fast Search and Find of Density Peaks". In: *Science*. DOI: 10.1126/science.1242072. URL: https://www.science.org/doi/abs/10.1126/science.1242072 (visited on 01/20/2022).

Bibliography II

Vedaldi, Andrea and Stefano Soatto (2008). "Quick Shift and Kernel Methods for Mode Seeking". In: *Computer Vision – ECCV 2008*. Ed. by David Forsyth, Philip Torr, and Andrew Zisserman. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, pp. 705–718. DOI: 10.1007/978-3-540-88693-8_52.