Clustering Techniques for On-Demand Transport Data: a case study

Pedro Afonso, Ana Alves
Pedro Afonso
a21240004@isec.pt[1]

Biography

- Carlos Pedro Marques Afonso
- 23 years old
- Graduated in Computer Engineering from the Polytechnic Institute of Coimbra, more specifically at the Instituto Superior de Engenharia de Coimbra
- Finishing the Master's degree in Informatics and Systems, software development branch
- Computer Engineer at The Municipality of Estarreja

Table of contents

Transport On Demand has a great impact on urban mobility

Motivations

- Reduce the number of vehicles needed
- Reduce the amount of pollution needed
- Decrease the difficulties in arranging parking
- Increase the flow of transit routes

Goals

- Analyse existing clustering algorithms, namely K-Means and DBSCAN
- Choose the best one to use in the case study
- Apply the algorithm to a sample dataset and evaluate it to define pickup zones that can help us improve transport routes

Overview of Related Work

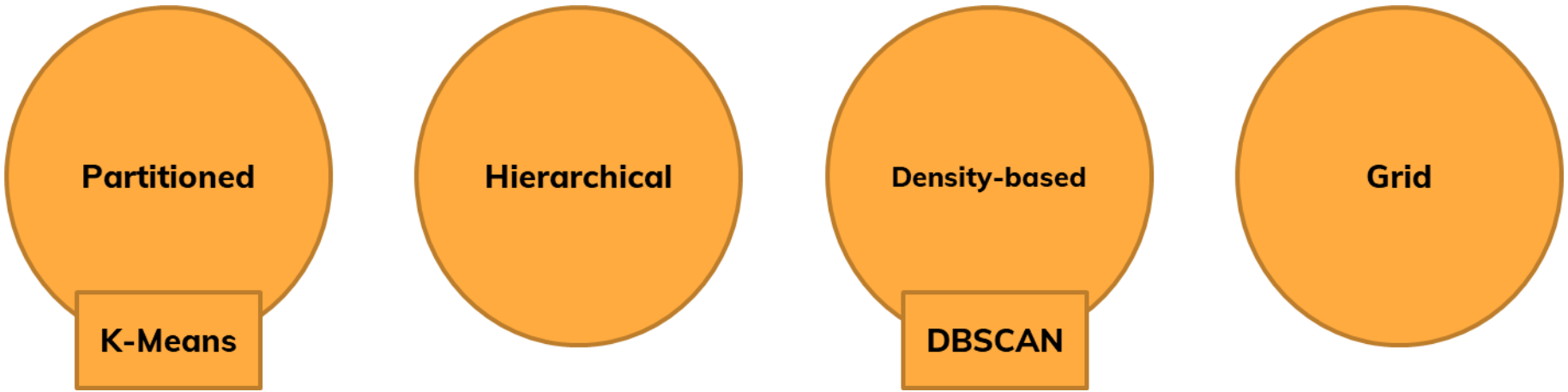
- DBSCAN was already used to group other types of geo-referenced data apart from transport requests

- It was found to be the most used for on-demand transport domain

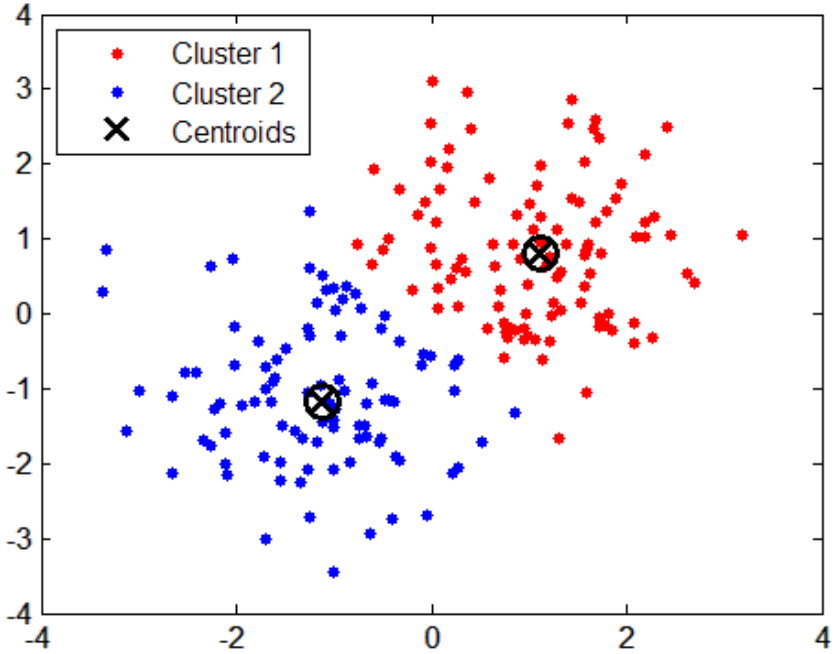
Grouping of people

- There are financial incentives and methods of road organizations to group as many people in each vehicle as:
    - **High-occupancy vehicle lanes**
    - **High-occupancy toll lanes**
    - **Slugging lines**

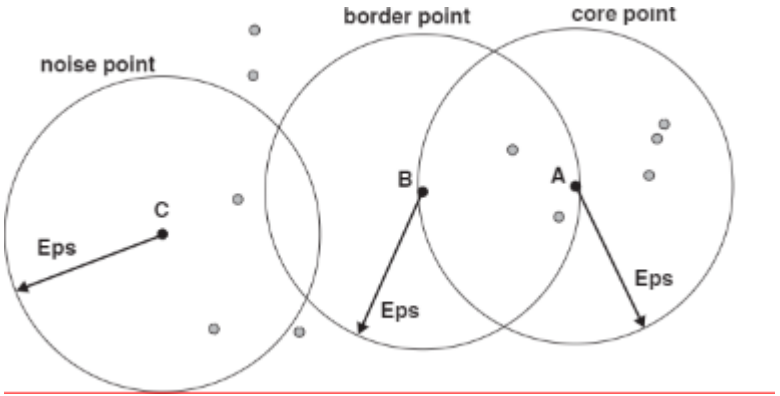
Clustering

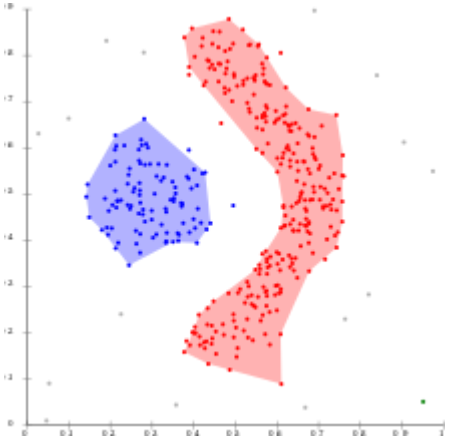

Partitioned
K-Means
Hierarchical
Density-based
DBSCAN
Grid

K-Means



DBSCAN





Technique comparison

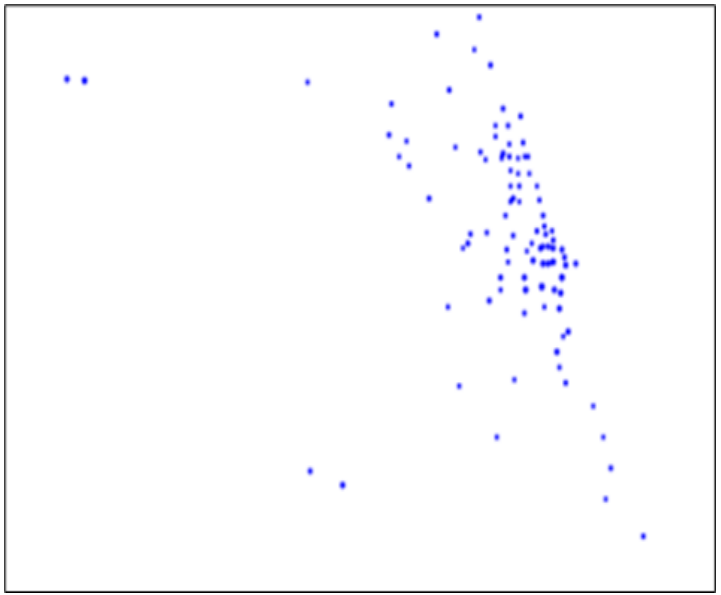
- Feature A: ability to identify clusters with random shapes
- Feature B: ability to identify clusters in datasets with high data volume
- Feature C: good performance in obtaining results
- Feature D: ability to deal with noise
- Feature E: parameterization/initial configuration of the algorithm
- Feature F: Handle numeric values

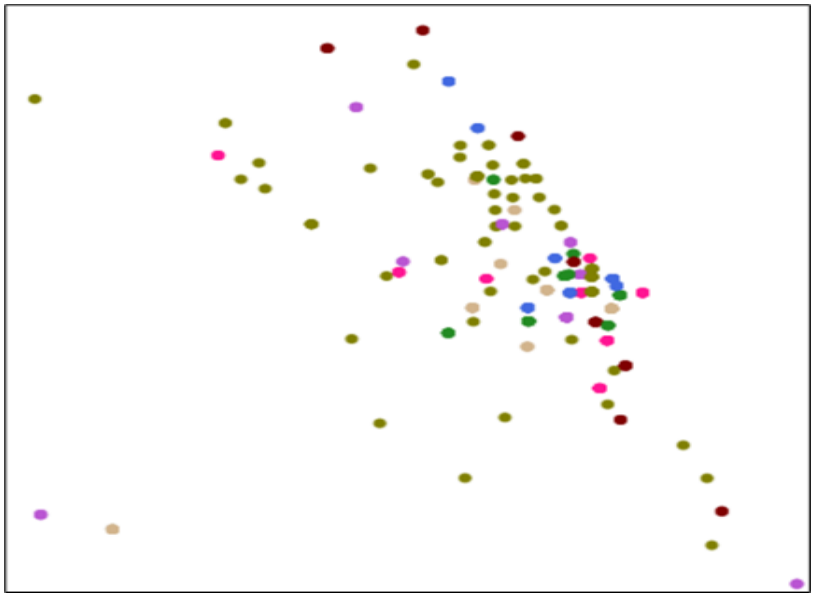
Quoted from the thesis of “Análise de dados e Machine Learning na Mobilidade Urbana” by João Pedro Fernandes Simões (https://comum.rcaap.pt/bitstream/10400.26/29858/1/Joao-Pedro-Fernandes-simoes.pdf)

Technique comparison

Algorithm	A	B	C	D	E
K-Means	X	X	X	X	✔
DBSCAN	✔	✔	✔	✔	✔

Case Study – Clustering pickup locations





https://github.com/pedroafonsoo/clustering_case_study_industrial_seminars/blob/master/dbscan_case_study_dbscan.ipynb

Conclusions

- DBSCAN algorithm is the most appropriate to aggregate transport requests compared to K-Means

- By applying the algorithm with a real data subset, we obtained the set of associated clusters that define pickup points, with a strong silhouette value, indicating quality in the result

Future work

- **Challenge: There should be a limit in the maximum number of elements in each cluster!**
    - Because in a vehicle there is a certain number of passengers and in a cluster, there can be more passengers than the maximum capacity that the vehicle allows

Future work

- Combine the DBSCAN clustering technique with Constrained K-Means

- Constrained K-Means restrict the capacity of the minimum and maximum number of points for each cluster and at the same time guarantee the optimization of the distance between the points

Thanks!