An Analysis on Clustering based on the Kruskal's Algorithm

Presented by Xiaobo Qian, Yangli Liu, Lu Wang, Jieqi Yang

4/29/2022

Agenda

- Project introduction
- The clustering approach based on Kruskal’s algorithm
- The K-means clustering
- The improvements to Kruskal’s algorithm clustering
- Conclusion

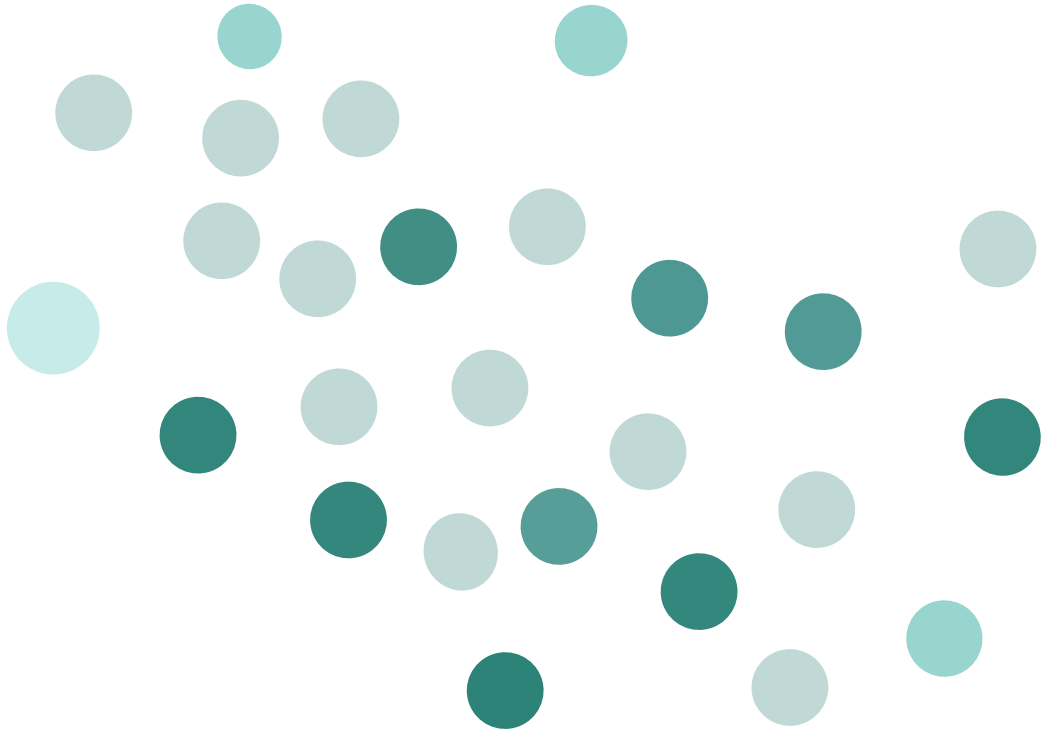
Agenda

- **Project introduction**
- The clustering approach based on Kruskal’s algorithm
- The K-means clustering
- The improvements to Kruskal’s algorithm clustering
- Conclusion

Project Background

Clustering is one of the major data analysis tools in the field of data mining and it is also a Machine Learning technique。

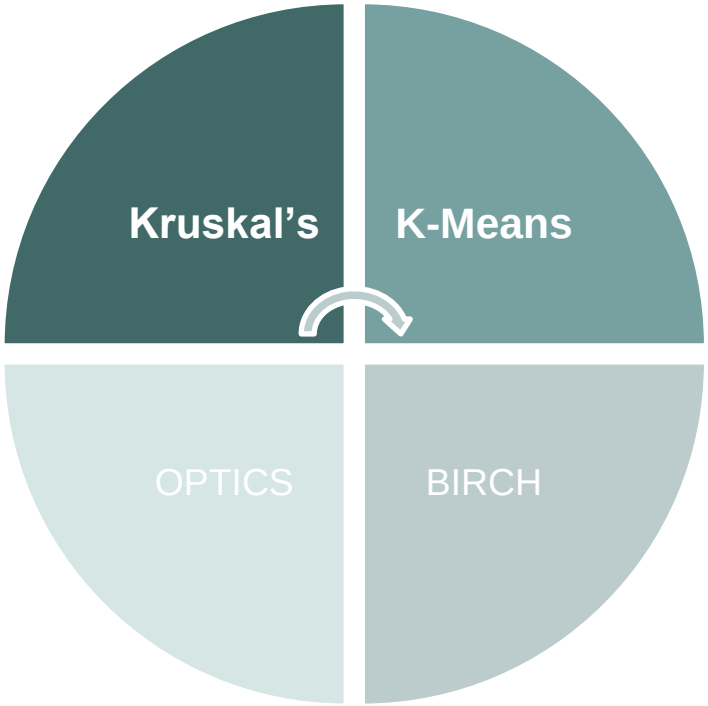
In practice, **clustering** helps identify the meaningfulness of data. (e.g. gene expression; customer segmentation)



Different Clustering Algorithms

Offer different approaches to the challenge of discovering natural groups in data

Have limitations in meeting their robustness and quality for clustering



Project Purposes

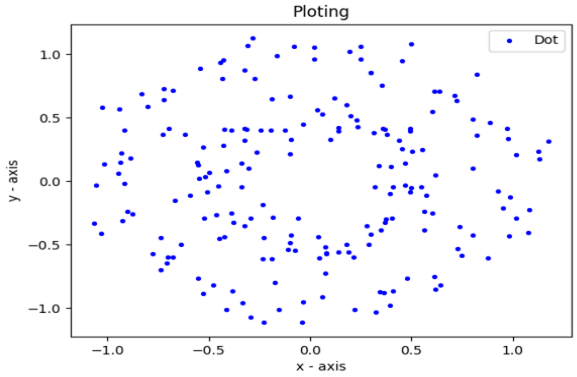
Kruskal's algorithm clustering

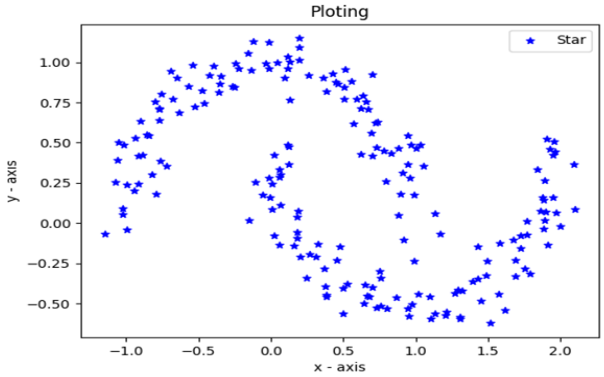
K-means clustering

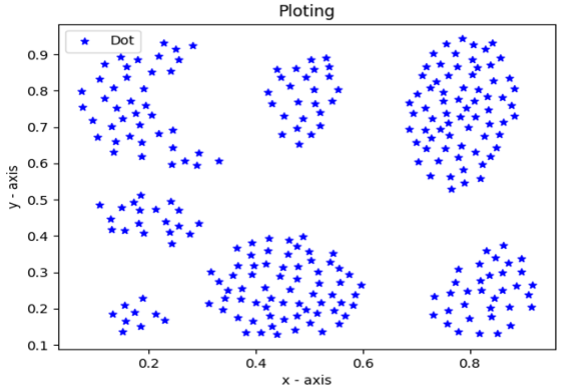
How we explore:

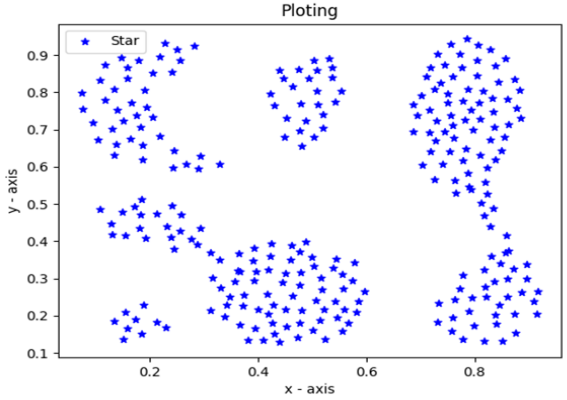
- **How the two algorithms work?**
- **Performance evaluation**
- **Improvement**

Dataset


Ploting
Dot
y - axis
x - axis


Ploting
Star
y - axis
x - axis


Ploting
Dot
y - axis
x - axis

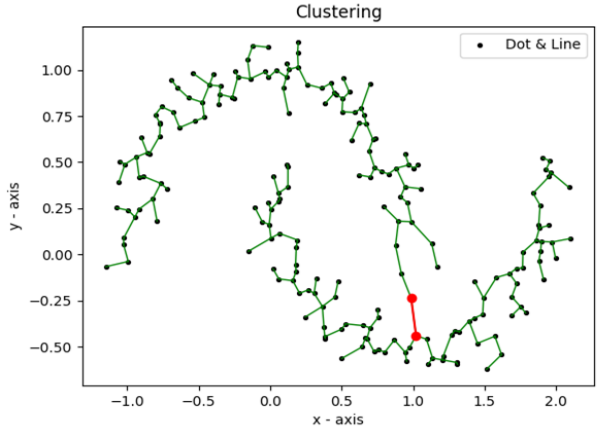

Ploting
Star
y - axis
x - axis

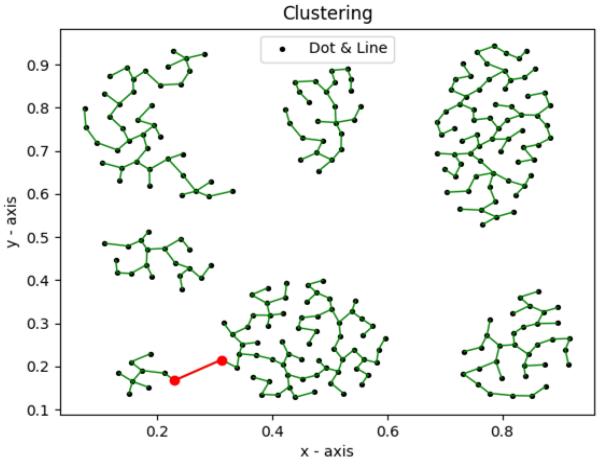
Agenda

- Project introduction
- **The clustering approach based on Kruskal’s algorithm**
- The K-means clustering
- The improvements to Kruskal’s algorithm clustering
- Conclusion

Kruskal’s Algorithm

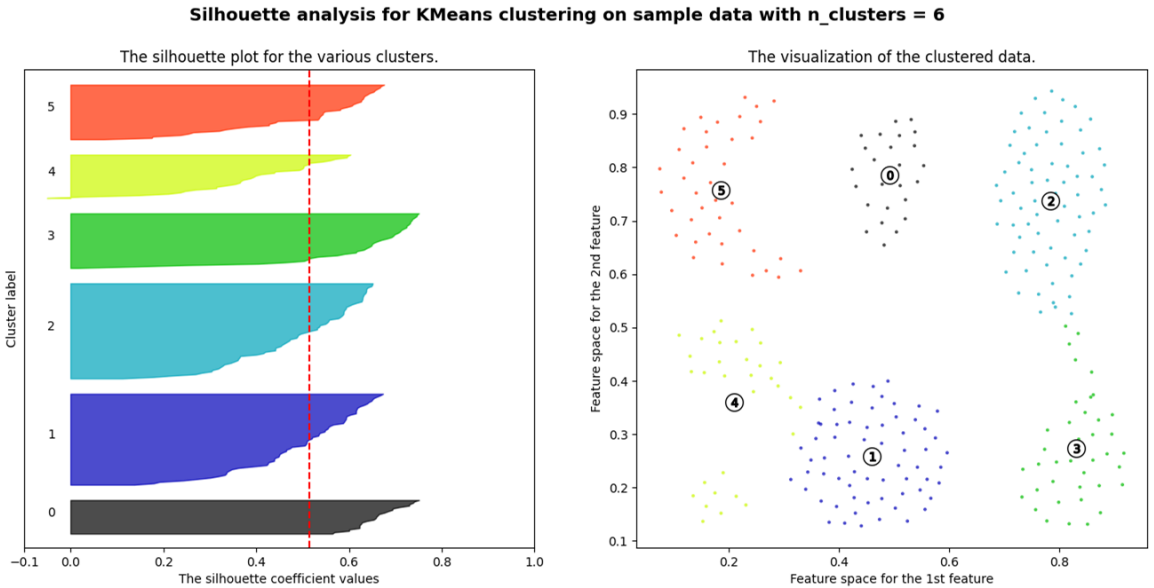
We have adapted what we learned from Kruskal’s algorithm into K-clustering algorithm. Each data point in K-clustering algorithm is a node with properties of parent and rank, and each node is a single cluster at the beginning. We maintain clusters as a set of connected components of a graph. After iteratively combining the clusters containing the two closest nodes by adding an edge between them, we will be able to draw the clusters in a forest-like way. The algorithm will stop when there are k clusters that k numbers are pre-selected by us. It will return the actual decimal numbers of the spacing of the clustering.





Picking K

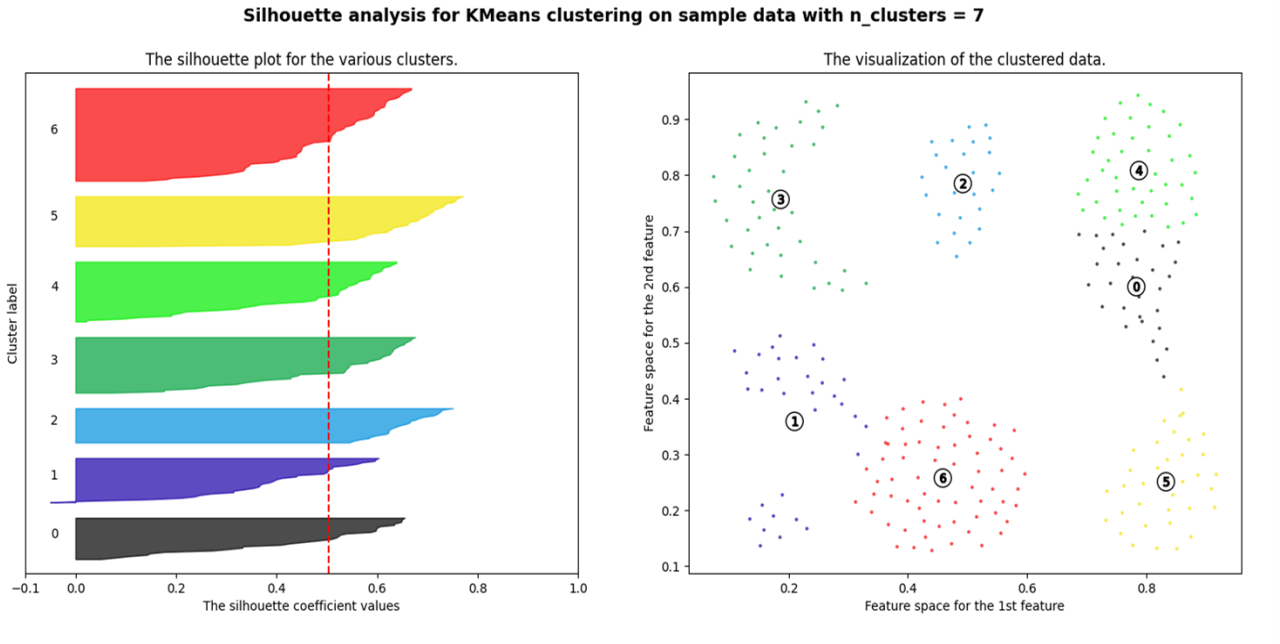
We did not mention this in our final report but we have tried different methods to find an optimal clusters numbers(value of k) for our K-clustering, like Elbow Method and Silhouette Method. Elbow Method is an empirical method to pick the value of k, where the average distance falls suddenly.



Silhouette Method

Silhouette method computes silhouette coefficients of each point and average it out for all the samples to get the silhouette score. This measure how much a point is similar to its own cluster compared to other clusters

$$s(i) = \frac{b(i) - a(i)}{max(b(i), a(i))}$$

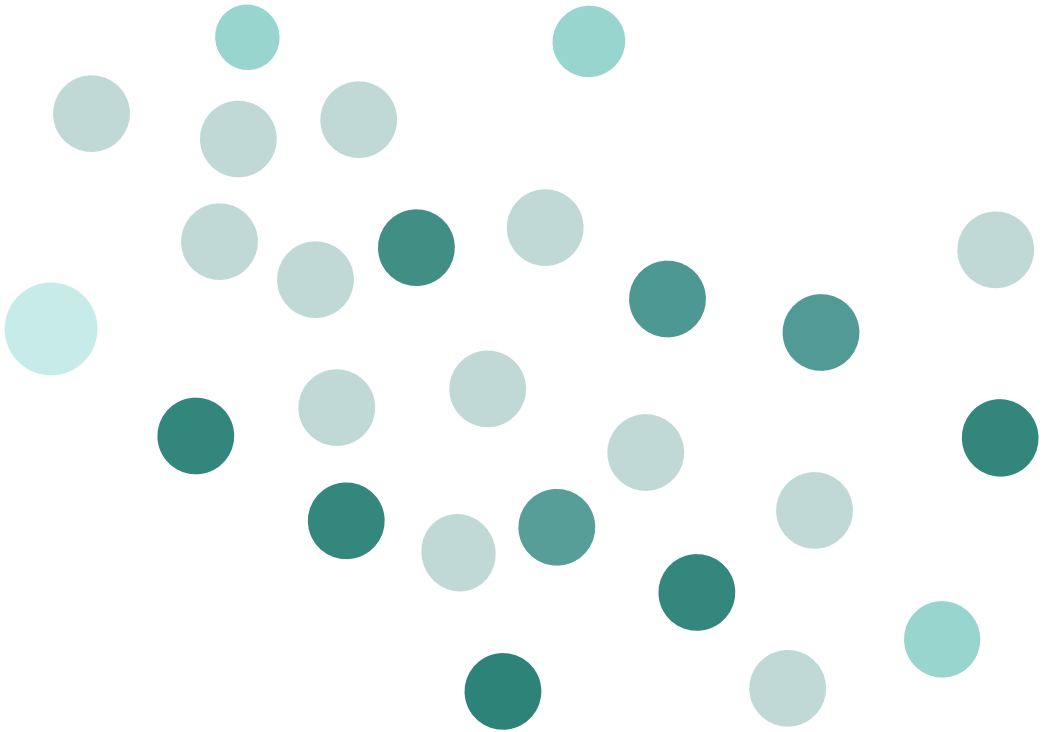


Agenda

- Project introduction
- The clustering approach based on Kruskal’s algorithm
- **The K-means clustering**
- The improvements to Kruskal’s algorithm clustering
- Conclusion

K-Means

K-means is an introductory algorithm to clustering techniques and it is widely used in unsupervised learning problems. K represents the number of clusters we are going to classify our data points into.



How does K-means work

K-means clustering repeatedly calculates and finds K centroids and group points with their nearest centroid.

The algorithm will stop when:

- Complete the maximum number of iterations;
- New formed centroids don’t change any more
- Clusters don’t change any more.

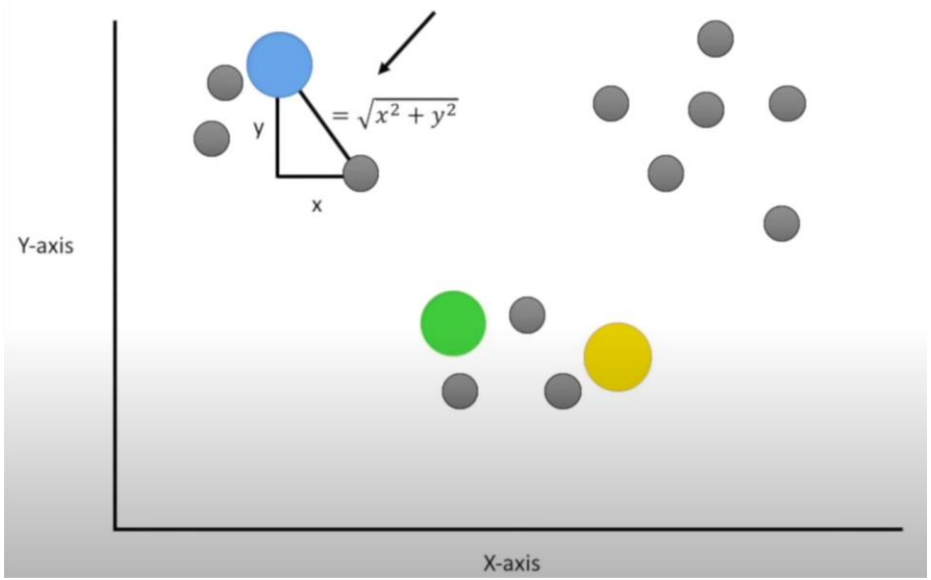
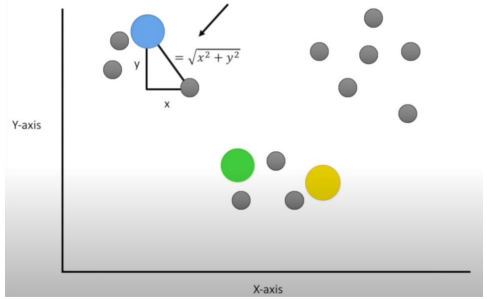


Image source: https://www.youtube.com/watch?v=4b5d3muPQmA

K-means pseudocode



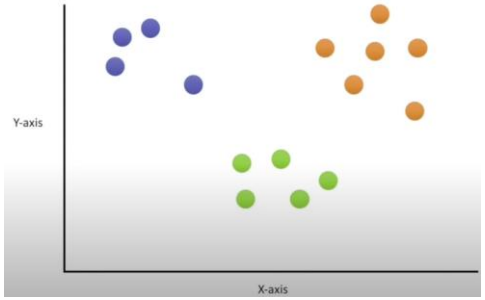


Image source:
https://www.youtube.com/watch?v=4b5d3muPQmA

Initialization

```
Import dataset D;

Specify the value of K and/or the maximum number of iteration if needed;

Choose the initial centroids c_1, c_2, ... c_k randomly;
```

Repeatedly find centroids

```
K-means(D, K):
   for each data point d_i:
      find its nearest centroid c_n among c_1, ... c_k;
      group d_i with c_n into a cluster;

   for each cluster j = 1..k
      Calculate and find its new centroid = mean of all points assigned to
cluster j

   Repeat the above two loops until convergence or until reach to the
maximum number of iterations
```

Performance Evaluation

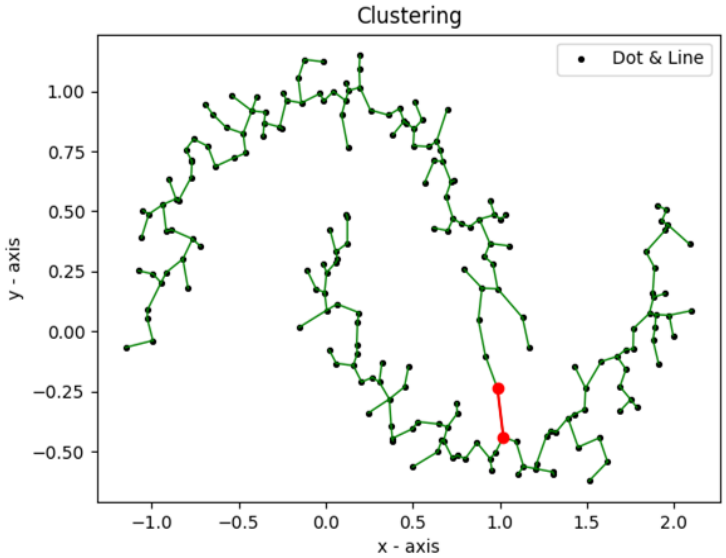
Kruskal

- No need to iterate multiple times
- Clusters are clearly separated

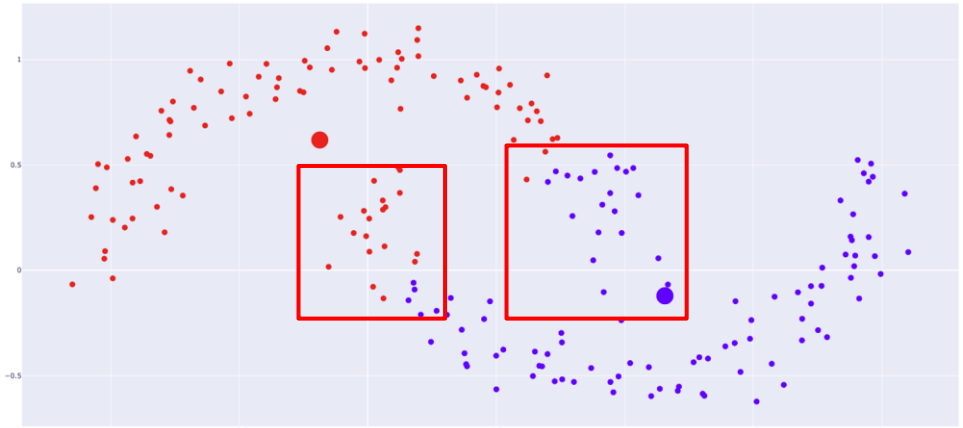
K-means

- Multiple times of iteration to find the centroids of the clusters
- Clusters were interspersed

Performance Evaluation

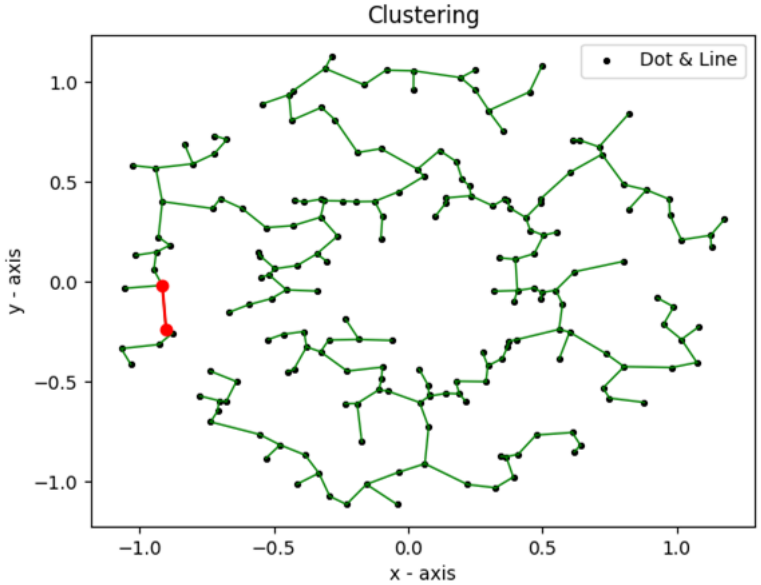


Two-moons by Kruskal

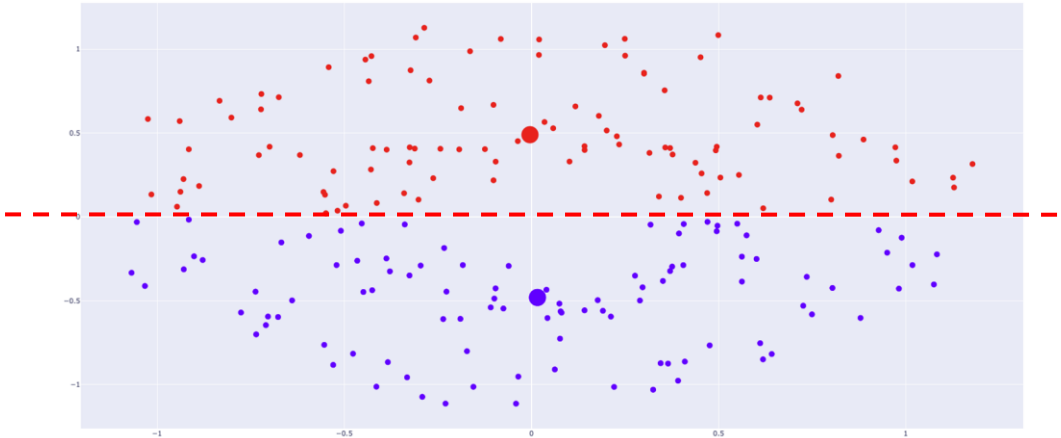


Two-moons by K-means

Performance Evaluation



Two-circles by Kruskal



Two-circles by K-means

Agenda

- Project introduction
- The clustering approach based on Kruskal’s algorithm
- The K-means clustering
- **The improvements to Kruskal’s algorithm clustering**
- Conclusion

Improvement

Improvements based on distance

Chebyshev distance is a metric defined on a vector space where the distance between two vectors is the greatest of their differences along any coordinate dimension.

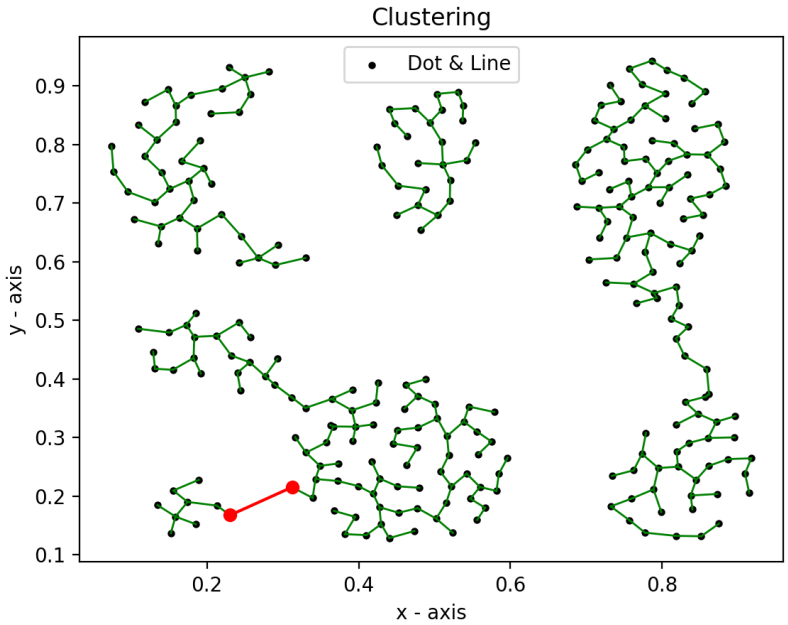
Manhattan distance is a distance metric between two points in a N dimensional vector space. It is the sum of the lengths of the projections of the line segment between the points onto the coordinate axes.

Improvement

Improvements based on distance

Mahalanobis distance is a measure of the distance between a point P and a distribution D. It is a multi-dimensional generalization of the idea of measuring how many standard deviations away P is from the mean of D.

$$d_{ij} = \left[(x_i - x_j)^T S^{-1} (x_i - x_j)\right]^{\frac{1}{2}}$$



Kruskal clustering based on Mahalanobis distance

Improvement

Improvements based on density

CFSFDP(Clustering by fast search and find of density peaks) is one way of finding “density peaks”. They can be deem as cluster centers.

$$density_i = \sum_{j \neq i} exp(\frac{-d_{ij}^2}{d_c^2})$$

$$distance_i = min(d_{ij}) \ (if \ \exists \ density_j > density_i)$$

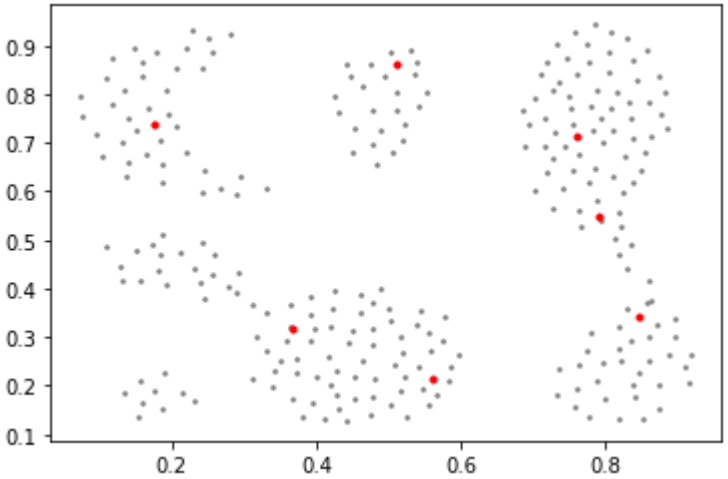
$$distance_i = max(d_{ij}) \ (if \ \forall \ j \neq i, \ density_j <= density_i)$$

$$score_i = density_i \times distance_i$$

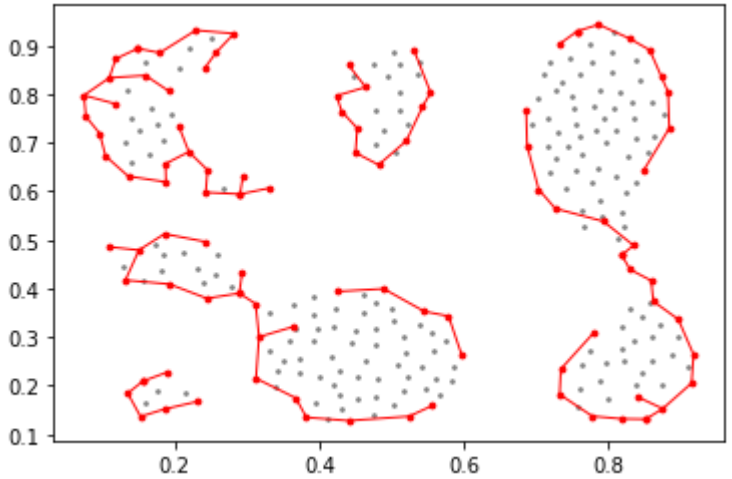
Improvement

Improvements based on density

CFSFDP(Clustering by fast search and find of density peaks) is one way of finding “density peaks”. They can be deem as cluster centers



“Imperfect Density Peaks” by naive CFSFDP



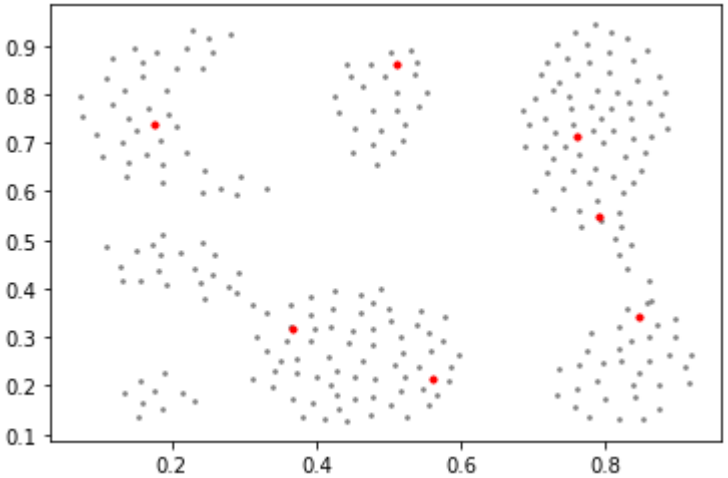
“Perfect Borders and Links” detection by naive CFSFDP

Improvement

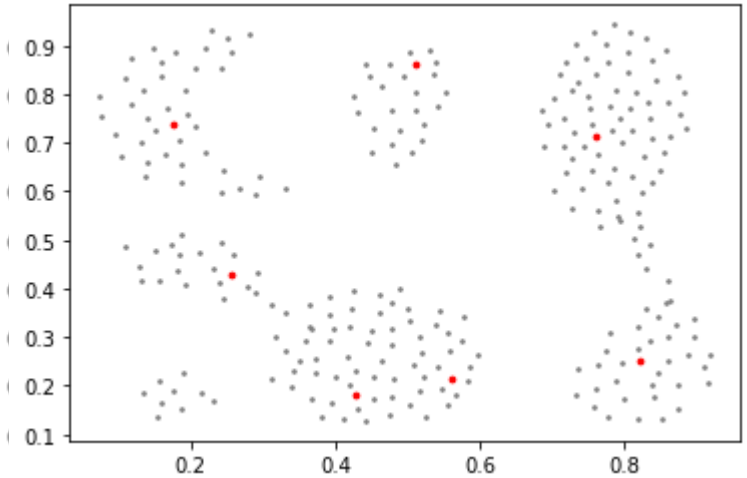
Improvements based on density

CFSFDP(Clustering by fast search and find of density peaks)

Fixing center positions of CFSFDP by elimination of low density vertices can help us evaluate the clustering results of Kruskal clustering.



"Imperfect Density Peaks" by naive CFSFDP



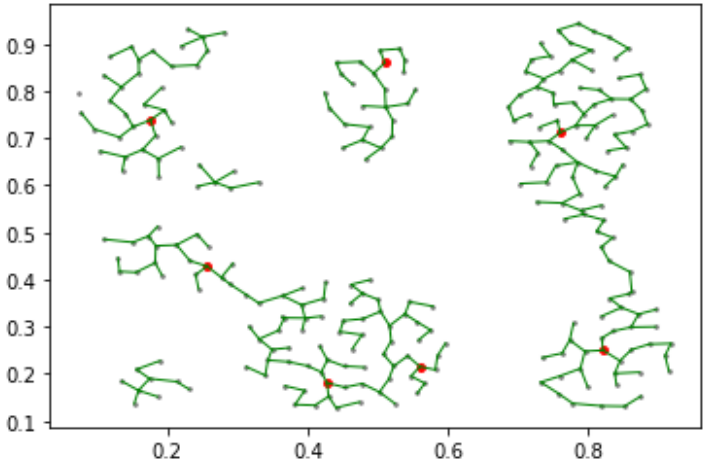
"Fixed Centers" by elimination of vertices

Improvement

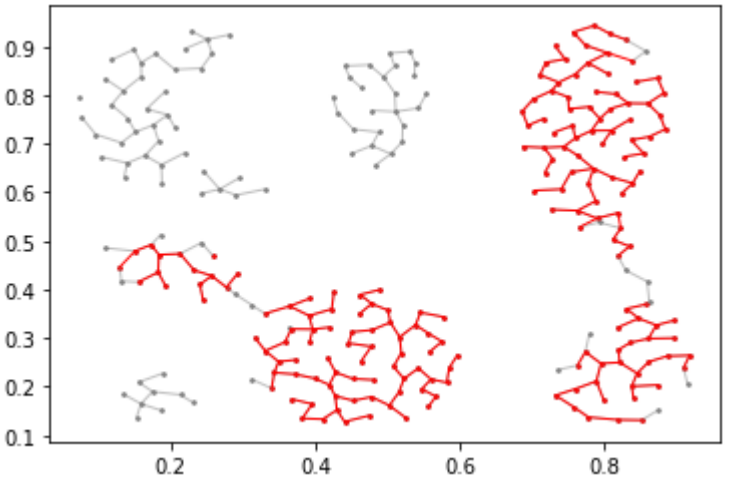
Improvements based on density

CFSFDP(Clustering by fast search and find of density peaks)

Clusters would be divided again if it contains more than 1 centers. Some vertices can be eliminated during the process.



Original Kruskal Clusters and “CFSFDP Centers”



Kruskal Clustering Improved by CFSFDP

Agenda

- Project introduction
- The clustering approach based on Kruskal’s algorithm
- The K-means clustering
- The improvements to Kruskal’s algorithm clustering
- **Conclusion**

Conclusions

- Kruskal’s clustering
- Comparison with K-means clustering
- Improvements to the Kruskal’s clustering

References

- Wang Peng, Wang Junyi, "A Clustering Algorithm Based on Find Density Peaks," Proceedings of 2017 the 7th International Workshop on Computer Science and Engineering, pp. 81-85, Beijing, 25-27 June, 2017.
- A. Rodriguez and A. Laio,”Clustering by fast search and find of density peaks,”Science,vol.344,pp.1492-1496,June.2014.
- P. K. Jana and A. Naik, "An efficient minimum spanning tree based clustering algorithm," 2009 Proceeding of International Conference on Methods and Models in Computer Science (ICM2CS), 2009, pp. 1-5, doi: 10.1109/ICM2CS.2009.5397966.
- https://medium.com/@mitanshupbhoot/comparative-applications-of-prims-and-kruskal-s-algorithm-in-real-life-scenarios-4aa0f92c7abc
- https://sites.google.com/site/dataclusteringalgorithms/clustering-algorithm-applications?authuser=0
- https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5397966
- https://www.researchgate.net/publication/11355740_Clustering_gene_expression_data_using_a_graph-theoretic_approach_An_application_of_minimum_spanning_trees
- https://docs.google.com/document/d/1wWkZLMgSnQsENh8NjrpcR-pdkRelwFr6SiuC31H48W4/edit#
- https://www.geeksforgeeks.org/k-means-clustering-introduction/
- http://datamining.rutgers.edu/publication/internalmeasures.pdf
- https://en.wikipedia.org/wiki/Mahalanobis_distance

Thank you!