SIMULTANEOUS CLUSTERING AND OUTLIER DETECTION

13C7032 - Nitin Mittal

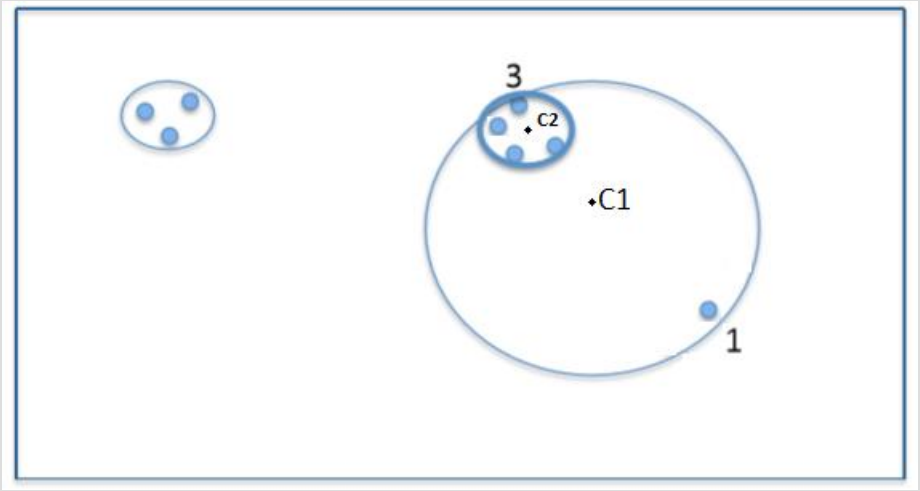
13C7037 - Pulkit Maloo

13C7054 - Tanay Kothari

PROBLEM DOMAIN

- Cluster centers are not accurately placed

- Setup Locations for -
- Telephone tower
- School
- Hospital



PROPOSED METHOD

- We propose a method for simultaneous outlier detection and clustering.
- Here we introduce a relative distance factor termed as “Outlier Factor“

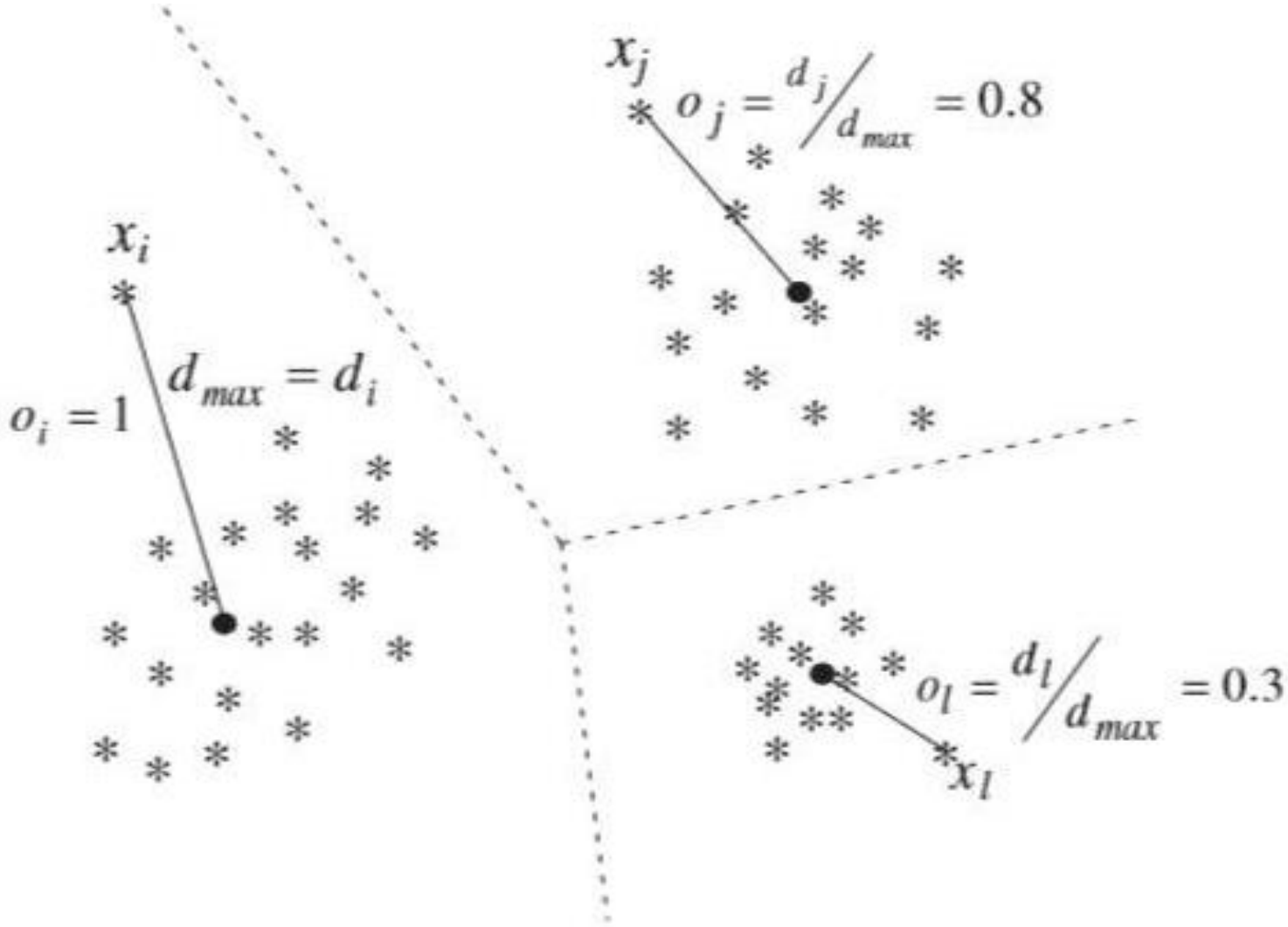
Outlier Factor –

- Defines outlying-ness of each data point in cluster
- Depends on distance of point from the cluster centroid
- It’s value will lie between 0 & 1

OUTLIER FACTOR - EVALUATION

- Calculating d_{max}
  - calculated for every cluster
  - dmax = distance of point farthest from centre.

$$\text{Outlier Factor} = \frac{\text{Distance of a Point from cluster centroid}}{d_{max}}$$


x_j
$o_j = d_j / d_{max} = 0.8$
x_i
$o_i = 1$
$d_{max} = d_i$
$o_l = d_l / d_{max} = 0.3$
x_l

THRESHOLD AND NO. OF ITERATIONS

- **Threshold**
  - Compared with outlier factor
  - Act as Bounding value for Outlier Factor
  - Magnitude lies between 0 and 1

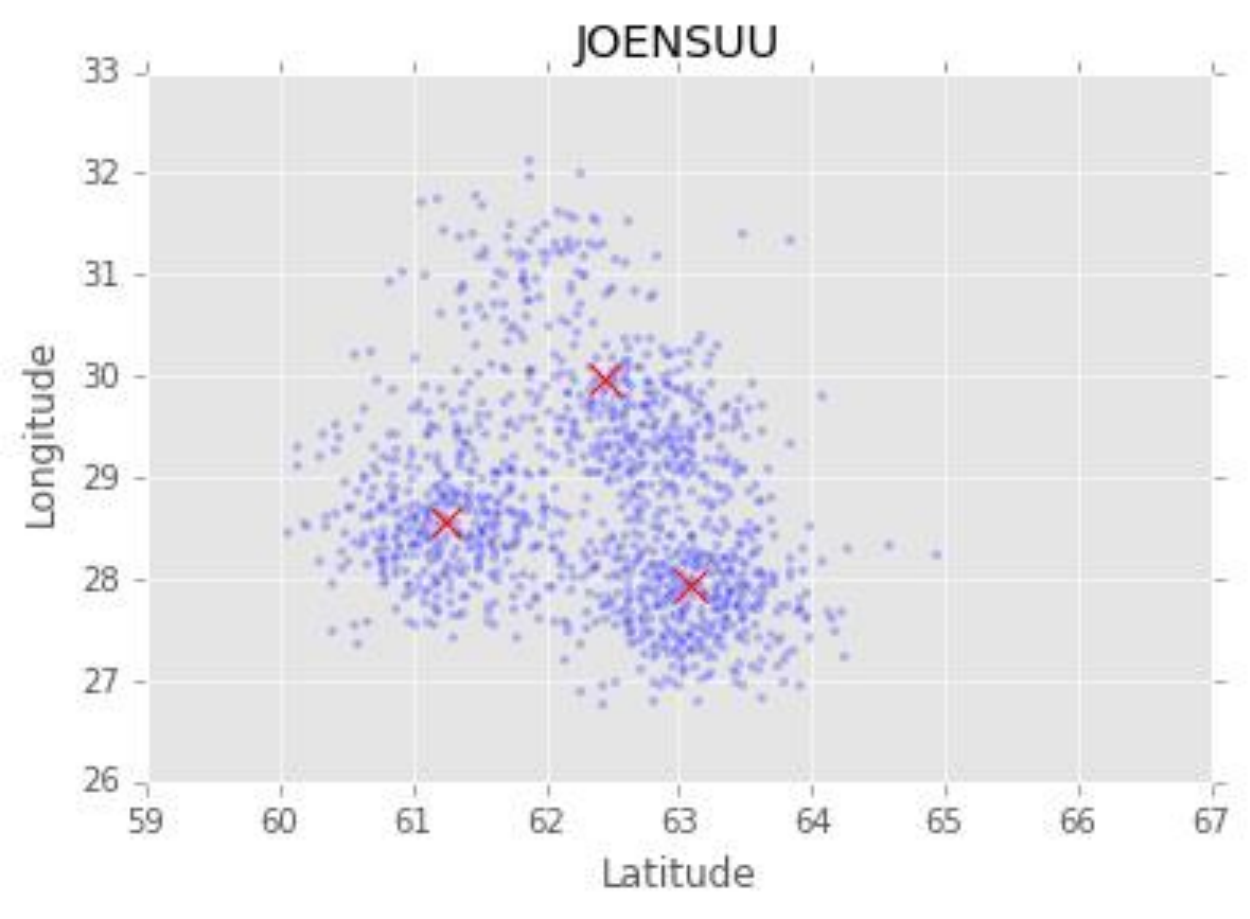
- **Iterations**
  - Determines number of times, user wants to compare outlier factor with threshold

CLUSTERING USING OUTLIER FACTOR

- 1. Run K-means until convergence, pick best Cluster Centroids C
- 2. For no. of iterations
- 2.1 For every Cluster Ci
    - a. Calculate outlier factor for each point
    - b. Remove all data points with Outlier Factor > Threshold, and update the dataset

2.2 Run K-Means over the updated dataset and update C

PRACTICAL OVERVIEW



Original Clustered Data –

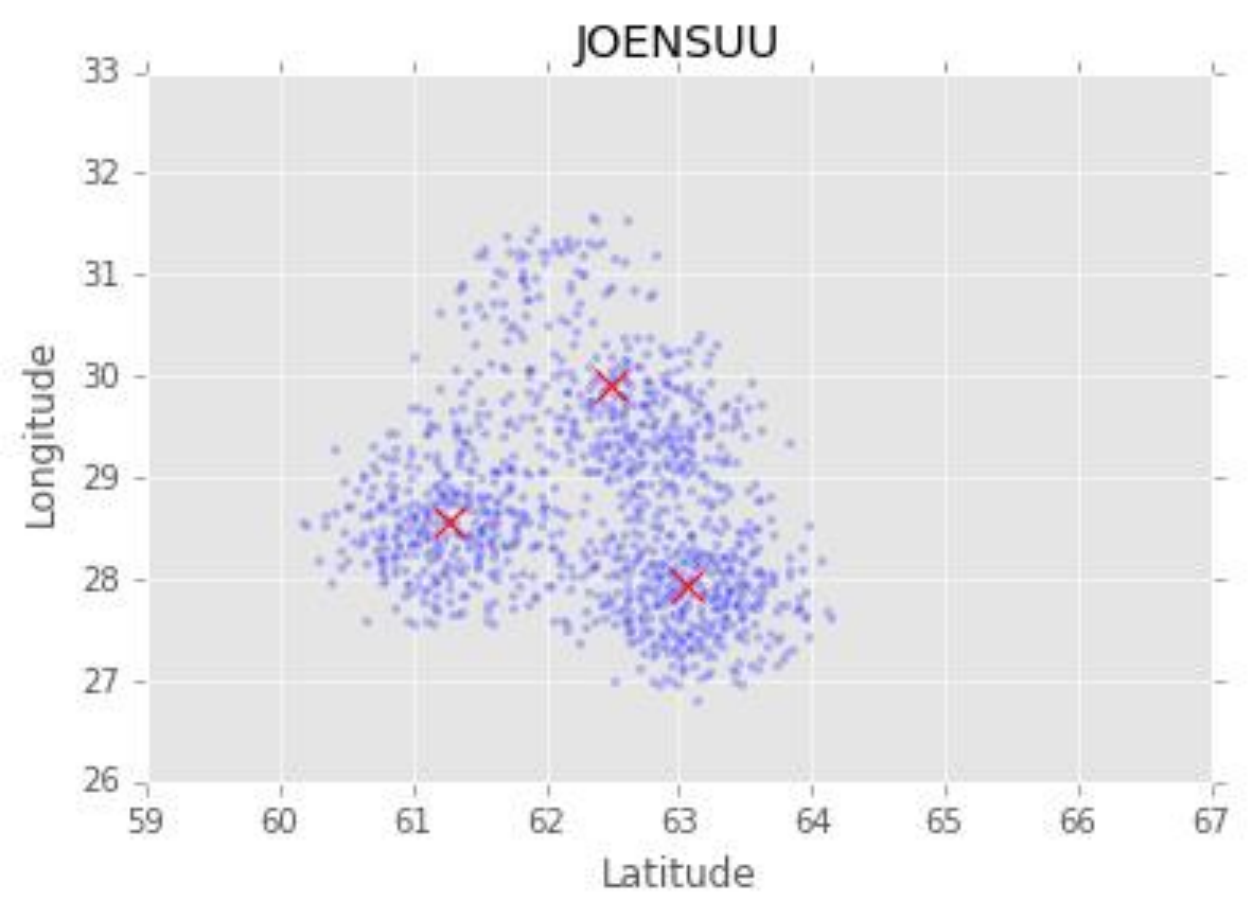
C1 = (61.24048782, 28.57907706)

C2 = (62.45291409, 29.98969893)

C3 = (63.09106467, 27.93085731)

Inertia = 725.097273378

PRACTICAL OVERVIEW



<u>After 5 iterations –</u>

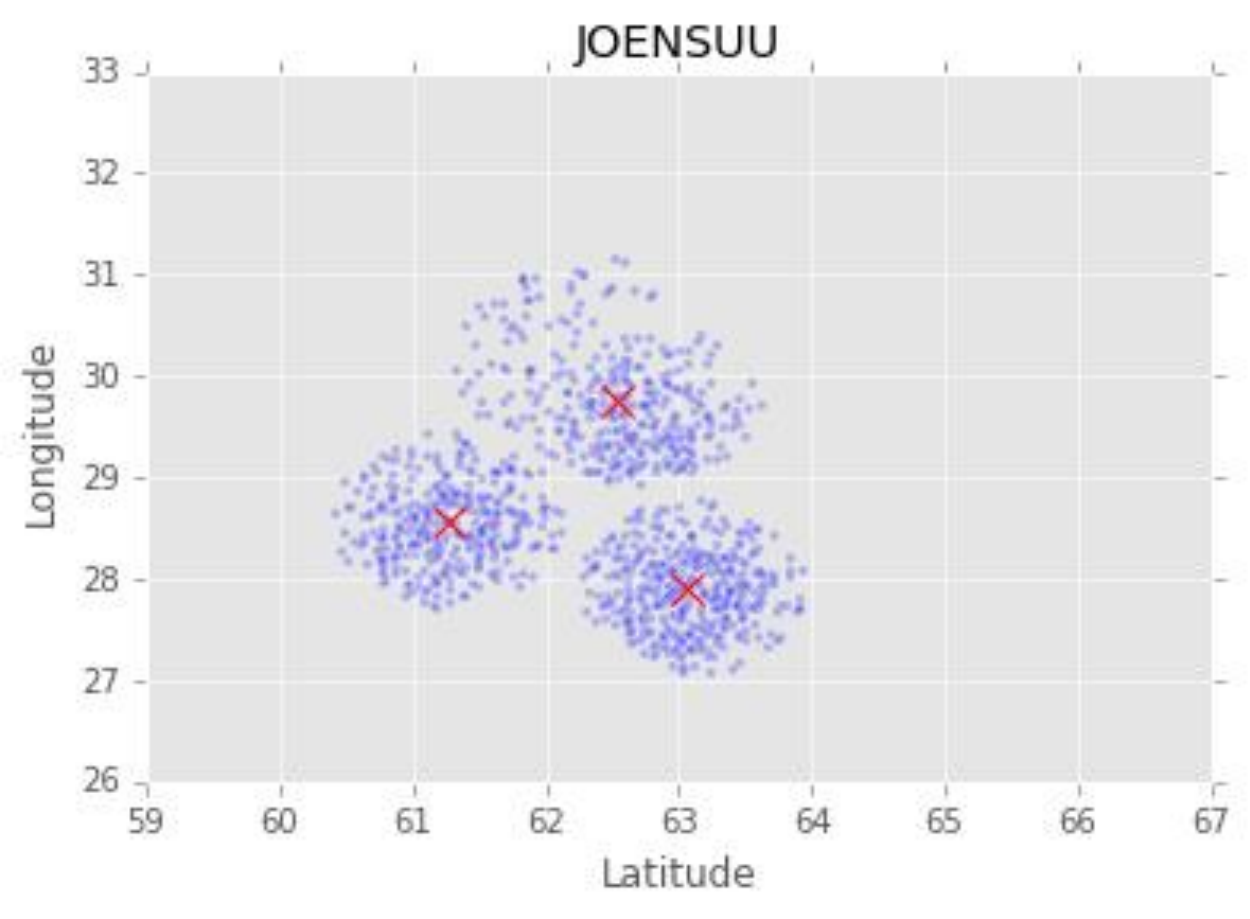
C1 = (61.27119879, 28.55349699)

C2 = (62.48589537, 29.90403648)

C3 = (63.07285294, 27.93593997)

Inertia = 739.293227297

PRACTICAL OVERVIEW



<u>After 10 iterations –</u>

C1 = (61.26480168, 28.55532924)

C2 = (62.54391747, 29.76397809)

C3 = (63.07163807, 27.8992316)

Inertia = 457.184761499

LIMITATIONS

- How to determine threshold T?
- T may differ from cluster to cluster
- Slow performance for very large datasets

LITERATURE SURVEY

1. Research issues on K-means Algorithm: An Experimental: Trial Using Matlab - Joaquín Pérez Ortega, Ma. Del Rocío Boone Rojas, María J. Somodevilla García

- Clustering Problem and the k-means Algorithm
- Sensitivity of K-Means towards outliers (noise).

2. Outlier Detection: A Clustering-Based Approach - Vijay Kumar, Sunil Kumar, Ajay Kumar Singh

- Absolute Distance between the Medoids and Point

LITERATURE SURVEY

3. k-means–: A unified approach to clustering and outlier detection - Sanjay Chawla, Aristides Gionisy

- Multivariate Outlier Detection

4. "Identification of Outliers". Chapman and Hall, London - D. Hawkins.

- Taxonomy of Outlier Detection Methods
- Multivariate Outlier Detection

5. Outlier Detection in Clustering - Svetlana Cherednichenko

- Distance-based approach

CONCLUSION

- Improving clustering by nullifying effects of outliers.
- Better positioning of Mobile Towers.
- Improves Efficiency.
- Have broad implementation.