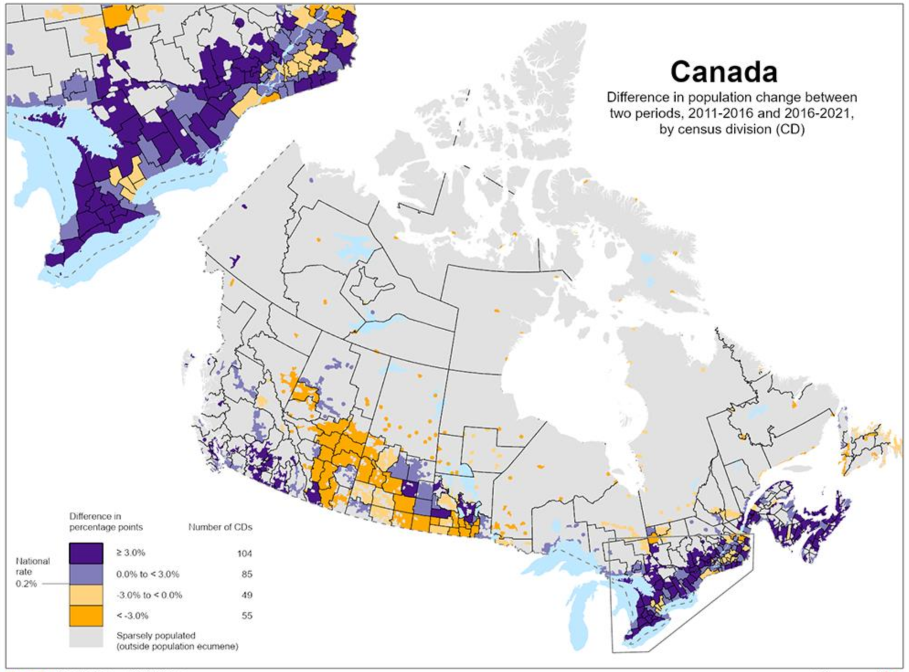
Data Clustering in Geospatial Informatics Applications



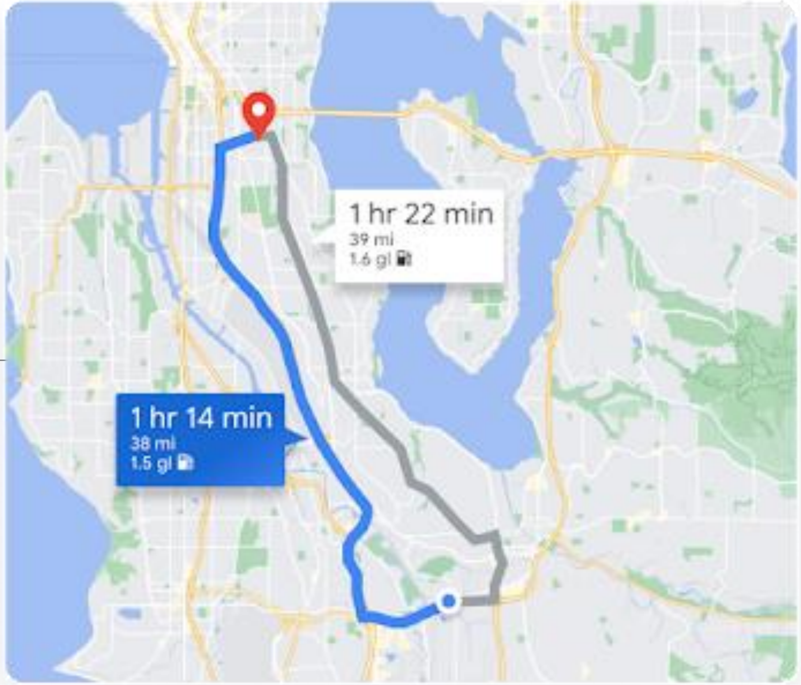
Canada
Difference in population change between two periods, 2011-2016 and 2016-2021, by census division (CD)
Difference in percentage points
Number of CDs
National rate 0.2%
≥ 3.0%
104
0.0% to < 3.0%
85
-3.0% to < 0.0%
49
< -3.0%
55
Sparsely populated (outside population ecumene)

Statistics Canada Statistique Canada

Source: Statistics Canada, Census of Population, 2011, 2016 and 2021.

Canada

ELIZABETH SWART



Introduction to Geospatial Data

Morais says that "80% of data is geospatial", although this point is arguable it identifies the prominence of geospatial data in our world.

Geospatial data is used in many contexts today: in disease tracking, climate change analysis, disaster response and transportation developments.

Geospatial big data is being used more often everyday, older algorithms are not equipped to handle this data and its unique challenges to process and analyze it. New techniques need to be developed to handle these new problems.

Geospatial Data Challenges

Big data has 5 V’s that can categorize the major difficulties when dealing with large datasets.

Volume

Variety

Velocity

Veracity

Visualization

Fuzzy C-Means Clustering

Algorithm

1. Select initial c centroids randomly, or using distance-based probabilities

2. Build a matrix that stores the membership value of each data point in association to each centroid

3. Recompute the means for each cluster using the weighted values of each data point

$$c_k = \frac{\sum_{x \in X} w_k(x)^h x}{\sum_{x \in X} w_k(x)^h}$$

4. Repeat steps 2 and 3 until each mean converges to a point

5. Assign data points to clusters where their weight is the highest

K-Medoid Clustering

Algorithm

1. Select the initial k medoids from the data set. One way is using the greedy algorithm to choose the k data points so that the points are more likely to belong to separate clusters

2. Build a matrix of each data points membership value to each cluster center.

3. Swap the cluster centers with all other data points and compute the distances produced by the new configuration. Keep the medoids that produce the smallest sum distance

4. The model can be represented as the following function

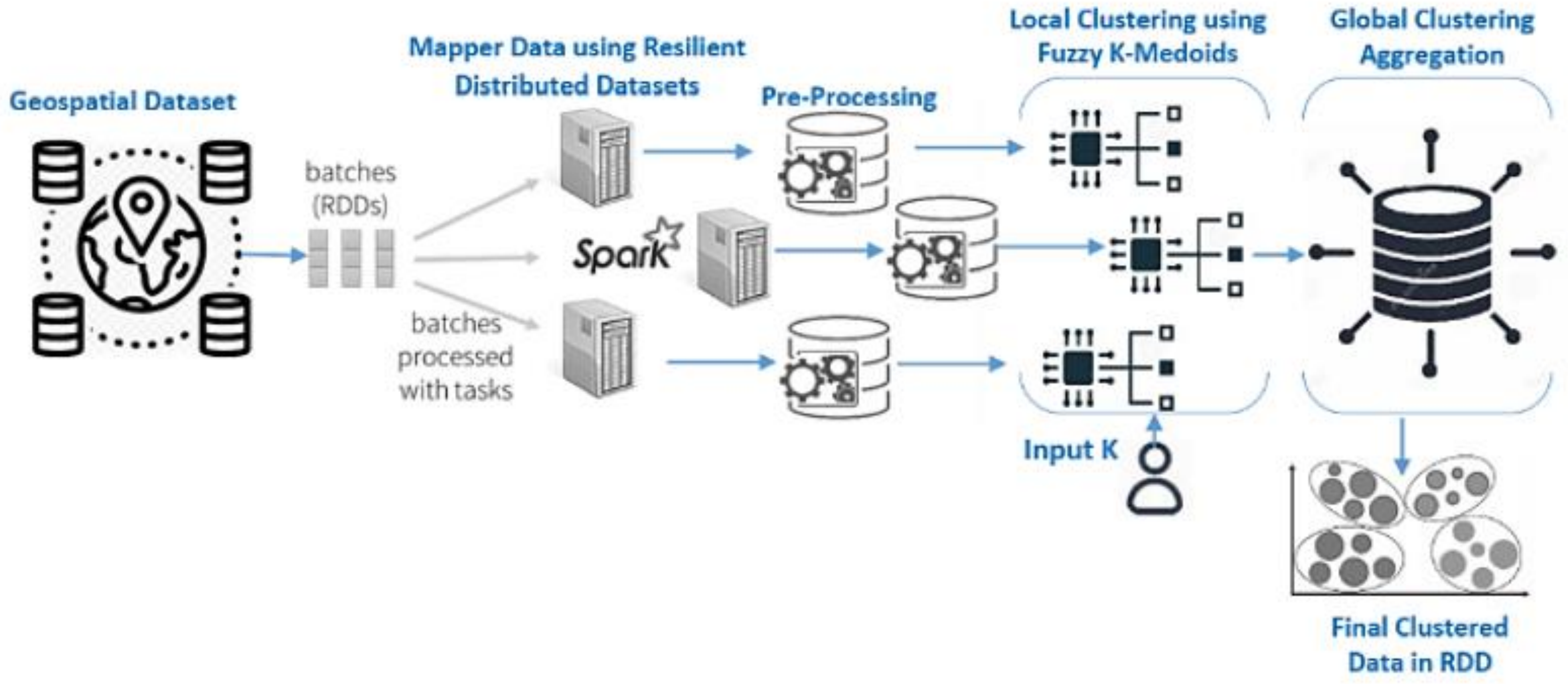
$$\min\ Z_{KM} = \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij} e_{ij}$$

$$\text{s.t. } \sum_{j=1}^{n} e_{ij} = 1,\ \forall\, i \in \{1, \ldots, n\},$$

$$e_{ij} \le e_{jj},\ \forall\, i, j \in \{1, \ldots, n\}, \sum_{j=1}^{n} e_{jj} = k,$$

$$e_{ij} \in \{0, 1\},\ \forall\, i, j \in \{1, \ldots, n\}.$$

Distributed k-Fuzzy Medoid Algorithm


Geospatial Dataset
batches (RDDs)
Mapper Data using Resilient Distributed Datasets
Spark
batches processed with tasks
Pre-Processing
Local Clustering using Fuzzy K-Medoids
Global Clustering Aggregation
Input K
Final Clustered Data in RDD

Local Clustering

The local clustering algorithm is an entropy-based variation of the fuzzy k-medoid algorithm.

This algorithm is used to reduce the dependency on initial k values and accounts for outliers and noisier data.

1. Select the initial k medoids

2. Cluster all data instances to the medoid with the minimum distance

3. Recalculate the medoids using the formulas:

$$\min Z_{FKM} = \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij} \left(e_{ij}\right)^{h} + \lambda \sum_{i=1}^{n} \sum_{j=1}^{n} e_{ij} \log(e_{ij})$$

$$e_{ij} = \frac{1}{\sum_{k=0}^{c} \left(\frac{d_{ij}}{d_{ik}}\right)^{\frac{1}{h-1}}}$$

4. Repeat steps 2 and 3 until convergence or maximum iterations is complete

Sampling the Clusters

Before the local clusters can be merged, the clusters are sampled to reduce the time complexity in the global clustering step. This process reduces the cost of comparing clusters from different nodes while maintaining the integrity of the cluster shape.

Types of data reduction: sampling, data compression, data discretization.

This algorithm uses a context aware reduction procedure that creates new cluster groups consisting of the medoid and the boundary points of the cluster. To find the boundary points the program iterates through all the points in a cluster, if a point is within a selected distance of any point in the cluster, then it is removed from the boundary group.

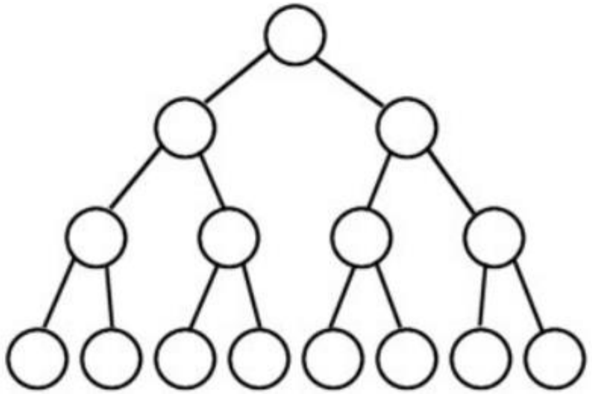
Global Clustering

The global clustering phase merges the local cluster groups into global clusters. This step is achieved by using a tree structure to merge the nodes together.

Starting at the lowest level, the program merges D Nodes together by selecting a leader node to be the processing unit and receive the information of all other nodes.

The leader node then merges overlapping clusters together to form new cluster groups.

The selection of a leader node is repeated with the information transmitted to the new leader and clustered by merging the clusters until a global clustering is built.



Improvements

In the paper proposing the distributed fuzzy-k algorithm, the cluster error, convergence time, and accuracy were tested against standard clustering techniques.

Cluster Error: The sum of all the distance between the objects in a cluster to the medoid for every cluster.

Clustering Accuracy: The percentage of data points that are correctly assigned to a cluster

Convergence Time: The convergence time is the number of time steps it takes to reach a global clustering of new data.

Challenges and Next Steps

The next steps to advance this algorithm is by reducing the time complexity. This can be done by determining the initial number of nodes needed in the distributed system to maximize productivity or by applying different soft clustering methods in the local clustering step to reduce the computing time.

Moving forward the next challenges faced with clustering big geospatial data is being able to build an algorithm that can analyze a stream of geospatial data for real time updates and queries.

References

Images

https://www150.statcan.gc.ca/n1/pub/92-195-x/2021001/other-autre/theme/theme-eng.htm

https://mapsplatform.google.com/maps-products/routes/

https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9706220

https://arxiv.org/pdf/1608.06861

https://www.mapsnworld.com/globe.html

Information

X. Deng et al, ‘‘*Geospatial big data: New paradigm of remote sensing applications*,’’ IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 12, no. 10, pp. 3841–3851, Oct. 2019.

S. Li, et al., *‘‘Geospatial big data handling theory and methods: A review and research challenges,’’* ISPRS J. Photogramm. Remote Sens., vol. 115, pp. 33–119, 2016.

M. M. Madbouly, et al. *"Clustering Big Data Based on Distributed Fuzzy K-Medoids: An Application to Geospatial Informatics."* IEEE Access, vol. 10, 2022, pp. 20926–20936.

N. Pinheiro, et al ‘‘*Convex fuzzy k-medoids clustering*,’’ Fuzzy Sets Syst., vol. 389, pp. 66–92, Jun. 2020

P. Praveen, et al, "*Big data environment for geospatial data analysis,*" 2016 International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2016, pp. 1-6.

A. Sabzi, et al, *‘‘An improved fuzzy k-medoids clustering algorithm with optimized number of clusters,’’* in Proc. 11th Int. Conf. Hybrid Intell. Syst. (HIS), Dec. 2011, pp. 206–210.

X. Yue, et a;, ‘‘*Parallel K-medoids++ spatial clustering algorithm based on MapReduce*,’’ 2016, arXiv:1608.06861.D.