Parallel Co-clustering for incomplete data

WU Zihan

September 29, 2023

Application

- **Incomplete data:** Clustering incomplete datasets involves accurately grouping samples that may be missing some attributes.
- **Existing Methods on Handling Incomplete Data:**
  - **Graph Representation Learning**[1]: Use graph representation learning to generate missing features.
  - **Adversarial Incomplete Multi-view Clustering**[2]: Infer missing data by discovering a common latent space.
- **Limitations:**
  - The incorporation of synthetic features can potentially introduce inaccuracies and skew the results.
  - Requiring substantial computational resources during the training phase.
- **Our solution:** Co-clustering can handle similarities between samples even if there are missing attributes

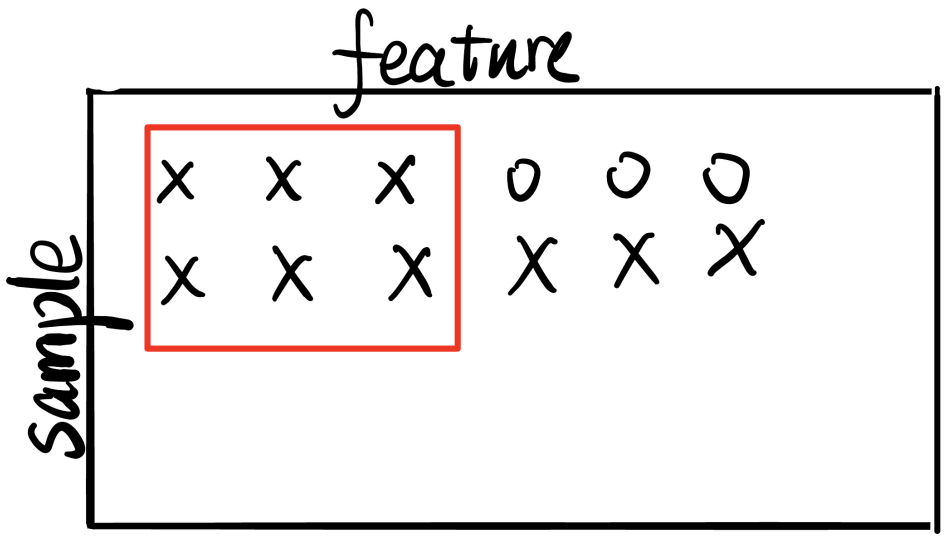
[1] You, J., Handling Missing Data with Graph Representation Learning, in: Advances in Neural Information Processing Systems, 2020.

[2] Xu, C., Adversarial Incomplete Multi-view Clustering, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, 2019.

Parallel Co-clustering for incomplete data

Key Contributions:

- **Co-clustering:**
  - Developed an enhanced technique for handling incomplete data.
  - Clarifies similarities between samples with missing attributes.
- **Parallelism:**
  - Optimized for efficiency and resource utilization.
  - Adaptable and scalable to diverse computing environments.
- **Probabilistic Model:**
  - Provides theoretical assurances of reliability and validity.
  - Ensures robustness and integrity of outcomes.



Research Findings and Observations

- **Big size simulation:** 100000×100000 matrix resulting good performance.
  - `[0.9921, 0.9832, 0.9941, 0.9937, 0.9715]`
- **Multi-Ellipses Expansion:**
  - Achieved 92.7% accuracy on pure ellipse images.
  - Struggled with accurate arc detection in noisy, multi-ellipse scenarios.
  - No clear application scenario identified.
- **Election and NLP:**
  - One research found on applying co-clustering to elections.
  - Identified potential for clustering text and extracting themes in NLP.
- **Food Nutrition and Market:**
  - Identified applications in recommender systems, clustering customers, and discerning preferences.
  - Potential to provide insights into consumer behavior and preferences in food nutrition and market domains.