Query-Efficient Correlation Clustering with Noisy Oracle

Yuko Kuroki Atsushi Miyauchi

Francesco Bonchi Wei Chen

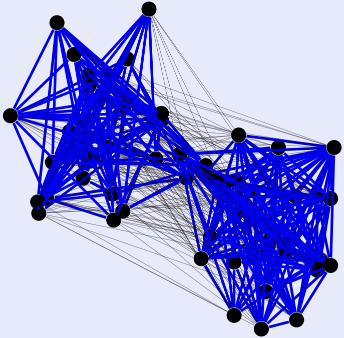
CENTAI

eurecat Centre Tecnològic de Catalunya

Microsoft Research

Correlation Clustering

Correlation Clustering (CC)



- A set $V = [n]$ of n objects and a pairwise similarity measure $\mathrm{s} : \binom{V}{2} \to [0, 1]$
- **Goal** of *Correlation Clustering* is to cluster the objects so that similar objects are put in the same cluster and dissimilar objects are put in different clusters.
- The objective is to minimize the following cost:

$$\mathrm{cost}_{\mathrm{s}}(\ell) = \sum_{\substack{(x,y) \in \binom{V}{2}, \\ \ell(x) = \ell(y)}} (1 - \mathrm{s}(x, y)) + \sum_{\substack{(x,y) \in \binom{V}{2}, \\ \ell(x) \neq \ell(y)}} \mathrm{s}(x, y),$$

where a clustering $\mathcal{C}$ can be represented as a function $\ell : V \to \mathbb{N}$.

Benefits and Approximability of Correlation Clustering

Benefits

- No need to know the number of optimal clustering (unlike k-means)
- Clustering based on pairwise judgments of similarity and dissimilarity (rather than quantitive distance information)

Approximation results for offline setting

- APX-hard! [Charikar, Guruswami, and Wirth 2003]
- Elegant $O(1)$-approximation algorithm, KwikCluster [Ailon, Charikar, and Newman 2008]

Query Efficient CC with Noisy Oracles

Research Question

We focus on the challenging scenario where

- (i) the underlying similarity measure is **initially unknown** and
- (ii) we can only query a **noisy oracle** that provides inaccurate evaluations of the weighted similarity $\mathrm{s}(x, y)$.

Goal: to devise clustering algorithms that perform as few queries on $\mathrm{s}(x, y)$ as possible to an oracle that returns noisy answers to $\mathrm{s}(x, y)$.

How We Leverage Bandit Theory

Pure Exploration of Multi-Armed Bandits (PE-MAB):

- Reduces the number of necessary queries by identifying the most informative pairs.
- Handles **noisy feedback** effectively through adaptive sampling.

Formulations as Pure Exploration of Bandits

At each round $t = 1, 2, \ldots$

- Learner will pull (i.e., query) one arm (i.e., pair of elements in V) from action space $E = \binom{V}{2}$ based on past observations.
- After pulling $e \in E$, the learner can observe the **random feedback** $X_t(e)$, which is independently sampled from an ***unknown*** distribution with mean $\mathrm{s}(e) \in [0, 1]$.

Fixed confidence setting

Given a confidence level $\delta \in (0, 1)$ and additive error $\epsilon > 0$, the learner aims to guarantee that $\mathrm{cost}_{\mathrm{s}}(\mathcal{C}_{\mathrm{out}}) \leq \alpha \cdot \mathrm{OPT}(\mathrm{s}) + \epsilon$ holds with probability at least $1 - \delta$. The evaluation metric of an algorithm is the sample complexity, i.e., the number of queries to the oracle the learner uses.

Fixed budget setting

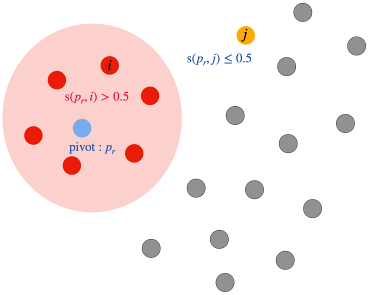
Given a querying budget T and additive error $\epsilon > 0$, the learner aims to maximize the probability that $\mathrm{cost}_{\mathrm{s}}(\mathcal{C}_{\mathrm{out}}) \leq \alpha \cdot \mathrm{OPT}(\mathrm{s}) + \epsilon$.

Key Ideas

Existing PE-CMAB's stopping conditions become no longer valid and the algorithm is not guaranteed to stop...!

Offline Algorithm Property

- Cluster the neighbors of a randomly selected pivot p_r at each phase $r = 1, \ldots, n$
- The **threshold** is $\mathrm{s}(x, y) > 0.5$
- Even without knowing $\mathrm{s}(x, y)$, we can design an online version of KwikCluster



KwikCluster

KwickCluster with Threshold Bandits

Identify High Similarity Pairs: Threshold bandits technique (e.g., [Kano et al. 2019])
Pivot Algorithm: Greedy procedure based on estimated high-similarity pairs

Output accuracy is guaranteed, even with small misidentifications for pairs with similarity close to 0.5.

Unbounded Samples: If pairs exist with $\mathrm{s}(x, y) = 0.5$, a naive threshold bandits algorithm may not guarantee termination.
Key Design for Efficient Sample Complexity: To prevent unbounded sample complexity, the threshold bandits step is designed to allow the misidentification of such pairs.

By accurately estimating the mean similarity between pairs (i.e., $\mathrm{s}(x, y) \in [0, 1]$), we can maintain an approximation guarantee of **5 in the offline (noise-free) setting**.

Theoretical Results

Approximation Guarantees and Sample Complexity (Informal)

The proposed algorithm guarantees a 5-approximate solution with additive error $\epsilon > 0$ w.p. at least $1 - \delta$. Sample complexity is:

$$T = O\left(\Sigma_{(x,y)\in E} \frac{1}{\tilde{\Delta}_{x,y}^2} \log \frac{1}{\tilde{\Delta}_{x,y}^2 \delta} + \frac{n^2}{\max\{\Delta_{\min}, \epsilon/n^2\}^2}\right),$$

where $\tilde{\Delta}_{x,y} \approx |\mathrm{s}(x,y) - 0.5|$ and $\Delta_{\min} := \min_{(x,y)\in E} |\mathrm{s}(x,y) - 0.5|$.

- This bound is much better than the uniform sampling method, which requires $T = O(\frac{n^6}{\epsilon^2} \log \frac{n}{\delta})$
- The significant term related to $\log \delta^{-1}$ is characterized by the gap $\Delta_{(x,y)}$, which represents the distance from 0.5.

Summary of More Results

Key Steps of KC-FB (Fixed Budget Setting)

- **Pivot-Based Exploration**: Randomly selects pivots to build clusters in phases.
- **Adaptive Query Strategy**: Allocates queries effectively by focusing on impactful pairs.

Experimental Highlights

- **Effectiveness**: Ours outperforms traditional uniform sampling methods, showing better clustering quality with fewer queries.
- **Scalability**: Demonstrates robust performance across different dataset sizes.

Conclusion

Correlation Clustering with Noisy Oracle

Approach: Bandit-Based Pure Exploration

- We introduced nove formulations rooted in PE-CMAB and algorithms whose sample complexity (number of queries) required to find a clustering whose cost is at most $5 \cdot \mathrm{OPT} + \epsilon$ with high probability.
- Our algorithms are the first examples of PE-CMAB for NP-hard offline problems.
- **Future work**: Deriving information-theoretic lower bounds of PE-CMAB for NP-hard offline problems, and investigating variants of correlation clustering or heteroscedastic noise scenarios.

Discover the Power of Bandit-Based Clustering!