Query-augmented Active Learning for Large-scale Clustering

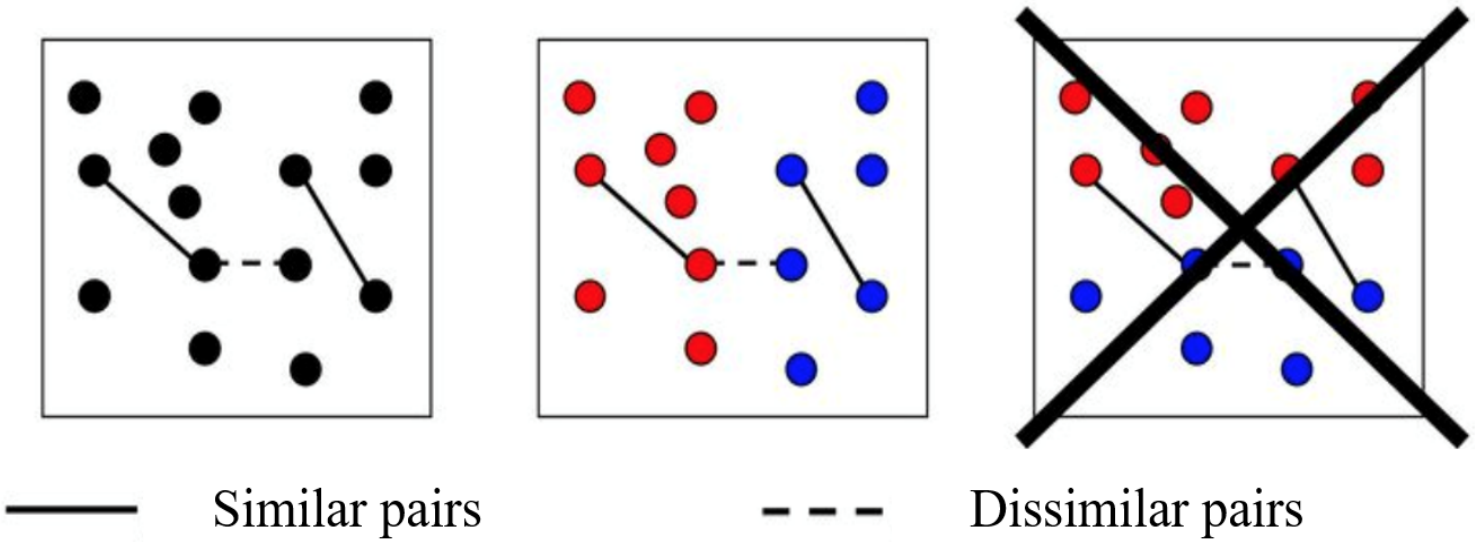
Yujia Deng[*], Yubai Yuan[†], Haoda Fu[‡] and Annie Qu[†]

[*]Department of Statistics, University of Illinois, Urbana-Champaign
[†] Department of Statistics, University of California, Irvine
[‡] Eli Lilly and Company

ICML Workshop, July 18, 2020

Problem formulation

- Cluster data with pairwise constraints



- Extract underlying feature space by learning a new metric
- Select the most informative unlabeled pairs actively

Contribution 1: Augmented metric learning

- Similar pairs tend to be close while dissimilar pairs far apart
- Metric learning with inferred pairwise constraints

$$\widehat{A} = \arg\min_{A} \frac{1}{|\mathcal{S}|} \sum_{(i,j)\in\mathcal{S}} \|x_i - x_j\|_A^2 + \frac{1}{|\widetilde{\mathcal{S}}|} \sum_{(i,j)\in\widetilde{\mathcal{S}}} w_{ij} \|x_i - x_j\|_A^2$$

$$\text{s.t.} \ \frac{1}{|\mathcal{D}|} \sum_{(i,j)\in\mathcal{D}} \|x_i - x_j\|_A + \frac{1}{|\widetilde{\mathcal{D}}|} \sum_{(i,j)\in\widetilde{\mathcal{D}}} w_{ij} \|x_i - x_j\|_A \geq 1, \quad A \succeq 0,$$

- A : metric, redefine distance, X: sample data, $\mathcal{S}$: similar pairs, $\mathcal{D}$: dissimilar pairs
- $\widetilde{\mathcal{S}}$: inferred similar pairs, $\widetilde{\mathcal{D}}$: inferred dissimilar pairs, w_{ij}: measures inference uncertainty
- **Utilize the information of unlabeled pairs in addition to labeled pairs to improve learning efficiency**

Contribution 2: Active query of data pairs

- Most informative pairs: labeling decreases uncertainty most
- Overall pairwise uncertainty

$$Q = -\sum_{i,j} \{p_{ij} \log_2 p_{ij} + (1 - p_{ij}) \log_2 (1 - p_{ij})\}$$

  - $p_{ij} = P(x_i \text{ and } x_j \text{ are similar})$
- Expected uncertainty of step $t + 1$:

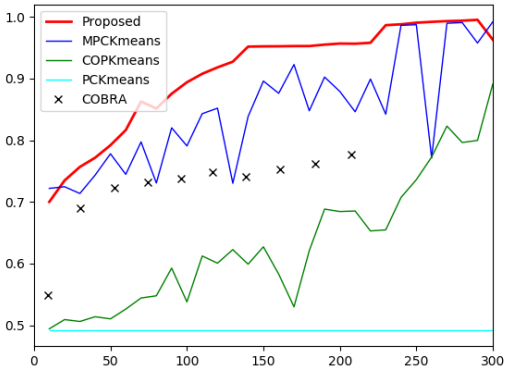
$$u^t(\boldsymbol{x}_i) = \sum_{m=1}^{L^t} r_{im}^t \widetilde{Q}_{-im}^{(t+1)}$$

  - $\widetilde{Q}_{-im}^{(t+1)}$: assume x_i belongs to m neighborhood with prob. r_{im}^t
- Active query to minimize the expected uncertainty of the next step

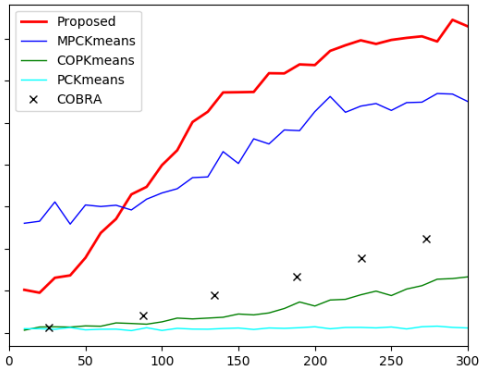
$$\boldsymbol{x}^* = \text{argmin}\, u^t(\boldsymbol{x}_i)$$

- **Instead of measuring uncertainty of step t, we compute the expected uncertainty of step $t + 1$**

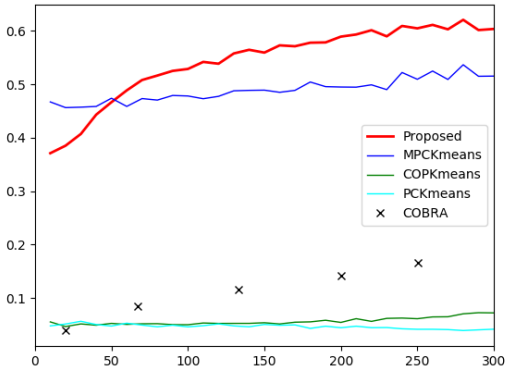
Results



(a) breast cancer



(b) MEU-Mobile



(c) urban land cover

Average ARI against number of constraints, compared with COPKmeans (Wagstaff et al., 2001), PCKMeans (Basu et al., 2004), MPCKMeans (Bilenko et al., 2004) and COBRA (Craenendonck et al., 2018). The first 3 methods are integrated with NPU (Xiong et al., 2014) query strategy

More in the paper:

- Metric aggregation through selective penalty, algorithms and theory