Crowdsourced Clustering via Active Querying

Yi Chen†, Ramya Korlakai Vinayak†, Babak Hassibi‡

† University of Wisconsin-Madison, ‡ Caltech

yi.chen@wisc.edu, ramya@ece.wisc.edu, hassibi@systems.caltech.edu

Crowdsourced Clustering

Problem Statement

Given n items, we want to cluster them into K disjoint clusters using noisy answers to pairwise queries from crowdsourced workers.

Related Work

- Yun and Proutiere [2] focused on the setting with **fixed** number of clusters of **large** sizes.
- Mazumdar and Saha [3] focused on the setting where the algorithm is aware of the error probability p.

Our Contribution

- Active clustering algorithm that does not rely on any unknown problem parameters like the number of clusters and workers' error rate.
- The algorithm is computationally efficient, simple to implement, and can recover clusters regardless of their sizes.

Problem Setup

- $\text{Query}(i,j) :=$ Are i,j from the same cluster?
- $X_{ij}(s) :=$ Answer of worker s to $\text{Query}(i,j)$
- $\text{cluster}(i) :=$ The cluster to which i belongs
- Assume $X_{ij}(s) \perp\!\!\!\perp X_{ij}(s')$ for $s \neq s'$

Two-coin Model for Worker Errors

When $\text{cluster}(i) = \text{cluster}(j)$, for all s,

$$X_{ij}(s) = \begin{cases} 1 & \text{with probability } p, \\ 0 & \text{with probability } 1-p. \end{cases}$$

When $\text{cluster}(i) \neq \text{cluster}(j)$, for all s,

$$X_{ij}(s) = \begin{cases} 1 & \text{with probability } q, \\ 0 & \text{with probability } 1-q. \end{cases}$$

We assume workers are better than random guessers, i.e. $1 \geq p > \frac{1}{2} > q \geq 0$.

Active Clustering Algorithm

- A randomly chosen item forms the first (singleton) cluster.
- Query a non-clustered item i with existing clusters.
- To decide if i belongs to $\mathit{cluster}(j)$
  - Item j picked randomly from $\mathit{cluster}(j)$,
  - Repeatedly make $\text{Query}(i,j)$ with different workers,
  - Until membership i can be established with **confidence**.
- Item i forms a new cluster itself if it is determined not to belong to any of the existing clusters.

The **confidence** is established by using the cumulative empirical average

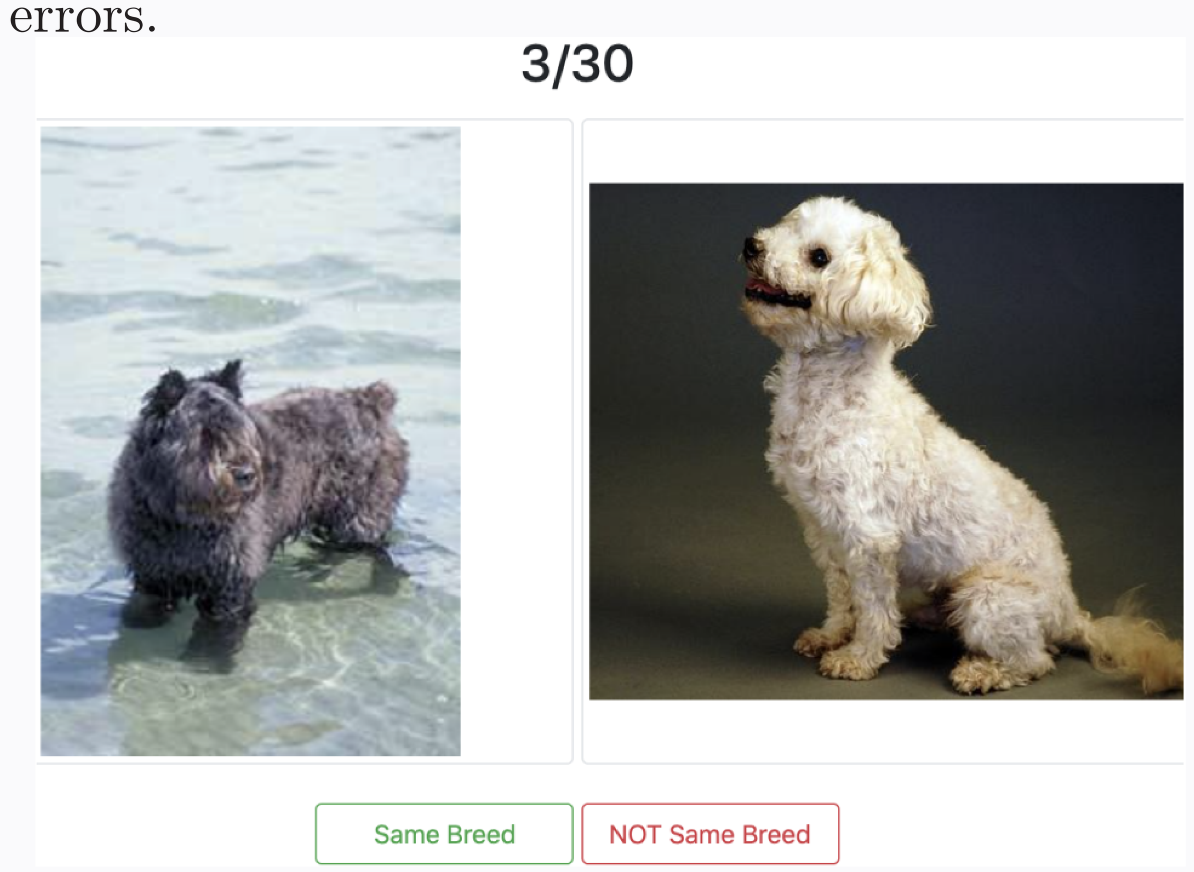
$$\bar{X}_{vu}(t) = \frac{t-1}{t}\bar{X}_{vu}(t-1)\frac{1}{t}X_{vu}(t),$$

and the confidence bound

$$\psi(t) = (1+\sqrt{\zeta})\sqrt{\frac{1+\zeta}{2t}\log\left(\frac{(1+\zeta)t}{\delta}\right)}.$$

- $\bar{X}_{uv}(t) - \psi(t) > \frac{1}{2} \implies v \in \text{cluster}(u)$
- $\bar{X}_{uv}(t) + \psi(t) < \frac{1}{2} \implies v \notin \text{cluster}(u)$

Note that δ and ζ are hyperparameters that are determined by your budget and your tolerance to errors.



Performance Guarantees

Theorem

Our algorithm succeeds in recovering all the clusters exactly with at most $\mathcal{O}(\frac{nK}{\Delta^2}\log n \log\frac{1}{\Delta})$, where $\Delta = \frac{1}{2}\min(p - \frac{1}{2}, \frac{1}{2} - q)$

Corollary

For any $\zeta \in (0,1)$, $c \geq 3$, $\delta = \frac{\delta'}{n^c} \in (0, \log(1+\zeta)/e)$, with probability at least $1 - \frac{1}{n}$, our algorithm succeeds in recovering all the clusters exactly and the total number of queries made is upper bounded by $\mathcal{O}(nK\frac{b_1}{\Delta^2}\log(\frac{n^c}{b_3\delta'}\log\frac{b_2}{\Delta}))$, where $b_1 = 3, b_2 = (1+\zeta)^2, b_3 = \frac{1}{(2(1+\sqrt{\zeta}))^3}$

Passive vs. Active

- **Active** querying succeeds **regardless of cluster sizes**.
- **Active** querying algorithm is **free of model parameters**.
- **Passive** querying followed by graph clustering can provide **good** clustering outcomes with fewer queries when **cluster sizes are large**.
- **Active** querying can pick up **more granular** differences within each cluster.

References

[1] Vinayak, Ramya Korlakai and Hassibi, Babak Crowdsourced Clustering: Querying Edges vs Triangles *Advances in Neural Information Processing Systems* pp. 328–332,(2016), NeurIPS

[2] Yun, Se-Young and Proutiere, Alexandre Community detection via random and adaptive sampling *Conference on learning theory*, pp. 138–175,(2014). PMLR.

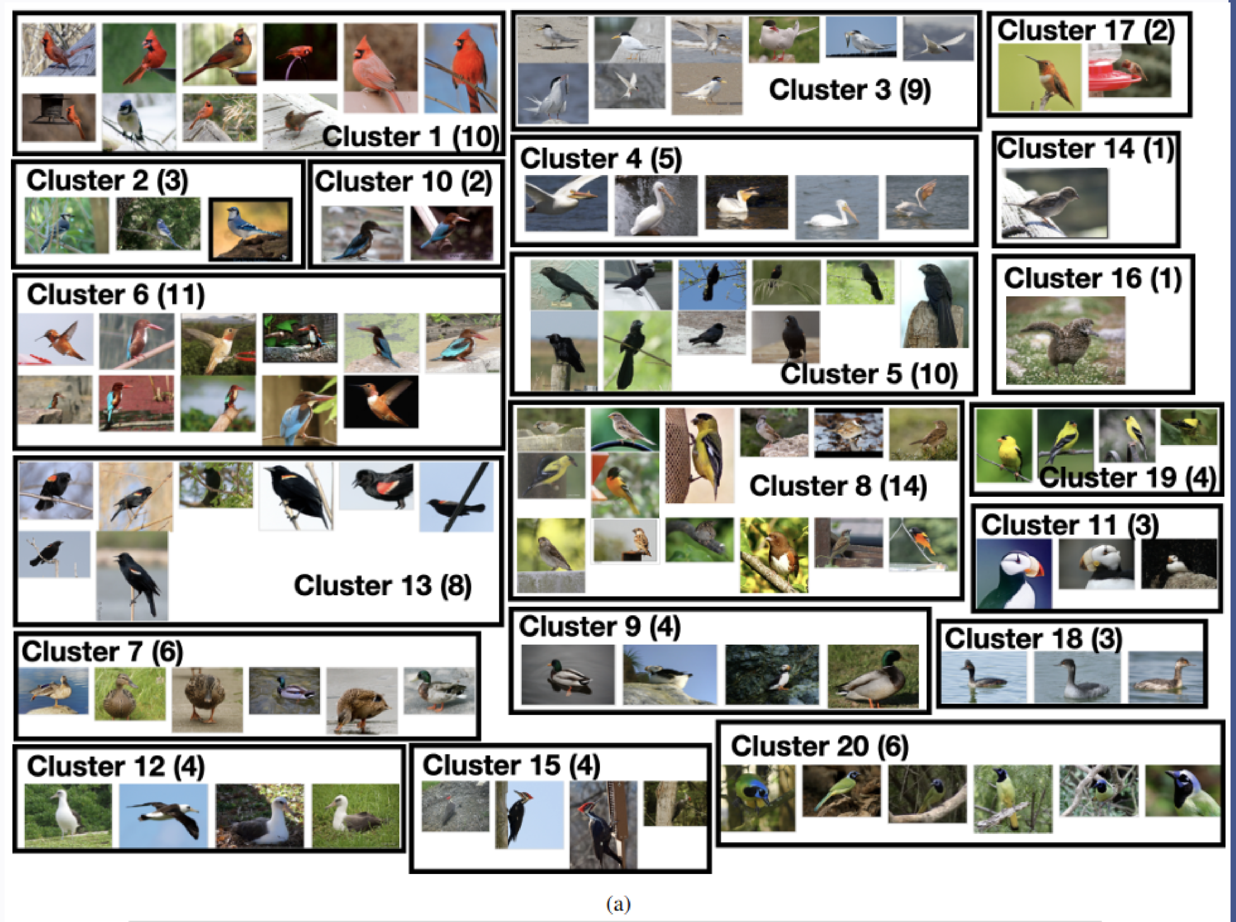
[3] Mazumdar, Arya and Saha, Barna Clustering with noisy queries *Advances in Neural Information Processing Systems* 30 (2017). NeurIPS.

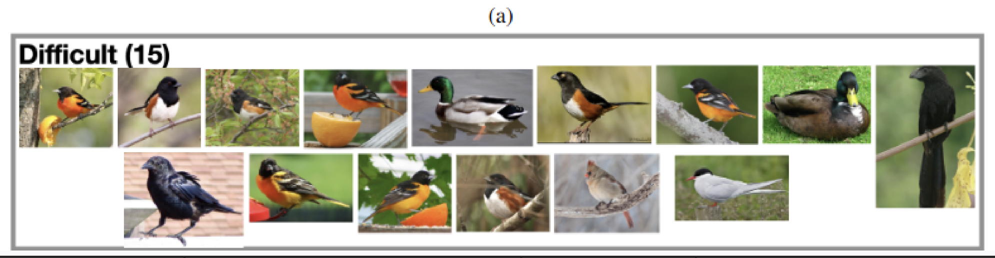
Experiment Results

Dogs3 dataset with large cluster size:

algorithm	pair error%	VI ↓	total queries
active	12.5%	1.85	43,572
passive	20%	**0.23**	**17,626**

Birds20 dataset with small cluster size:





algorithm	pair error%	VI ↓	total queries
active	1.69%	**0.88**	**15,160**
passive	18.4%	1.64	15,162