Correlation Clustering with Active Learning of Pairwise Similarities

Linus Aronsson, Chalmers University of Technology
Department of Data Science and AI

Abstract

Correlation clustering is a well-known unsupervised learning setting that deals with positive and negative pairwise similarities. In this paper, we study the case where the pairwise similarities are not given in advance and must be queried in a cost-efficient way. Thereby, we develop a generic active learning framework for this task that benefits from several advantages, e.g., flexibility in the type of feedback that a user/annotator can provide, adaptation to any correlation clustering algorithm and query strategy, and robustness to noise. In addition, we propose and analyze a number of novel query strategies suited to this setting. We demonstrate the effectiveness of our framework and the proposed query strategies via several experimental studies.

Background

Clustering constitutes an important unsupervised learning task for which several methods have been proposed in different settings. **Correlation clustering** is a well-known clustering problem that is particularly useful when both similarity and dissimilarity judgments are available between a set of objects. In this setting, a dissimilarity is represented by a negative relation between the respective pair of objects, and thus correlation clustering deals with clustering of objects where their pairwise similarities can be positive or negative numbers. The goal of correlation clustering is to find a clustering that minimizes disagreemnts (absolute value of the sum of negative similarities within a cluster plus the sum of positive similarities across clusters).

Problem Setting

In correlation clustering, one assumes access to pairwise similarities between all objects. However, as discussed in [1] computing such pairwise similarities can be computationally demanding, or they might not even be given a priori and must be queried from a costly oracle (e.g., a human expert). We employ active learning (AL) to query the most informative pairwise similarities in order to recover the ground-truth clustering with a minimal number of queries. The AL procedure consist of four steps:

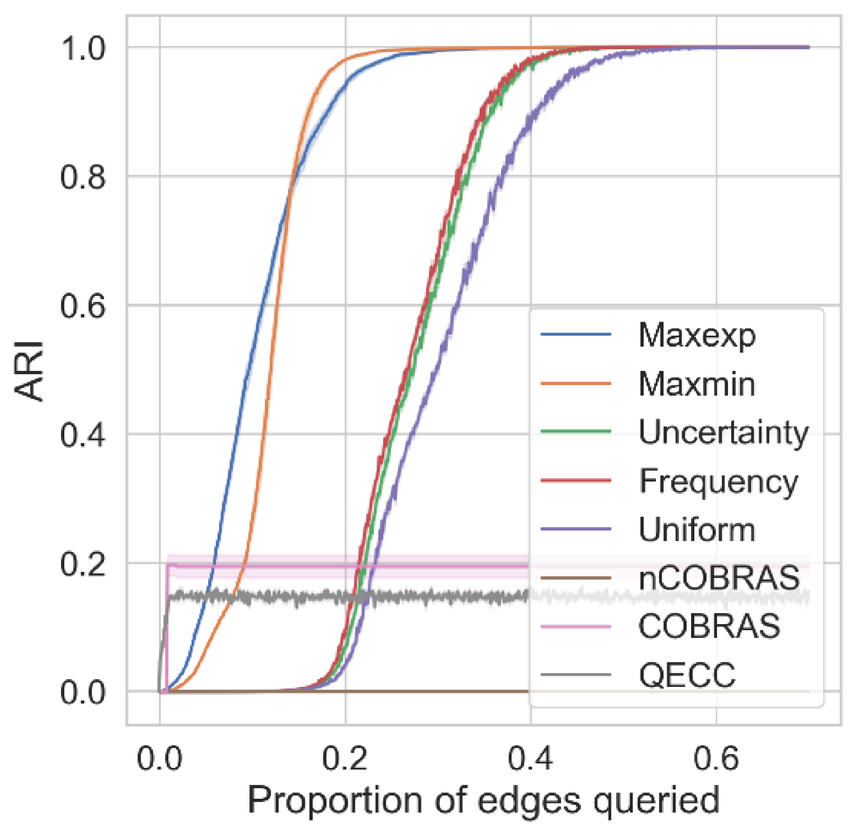
1. Initialize the similarities (e.g., randomly or based on prior information).
2. Run a correlation clustering algorithm given the current similarities, producing a clustering solution.
3. Query the oracle for the B most informative pairwise similarities.
4. Update the similarity matrix based on response from oracle.

Contributions

- Previous work are restricted to similarities in {-1, +1}. Instead, our methods can handle similarities that can be any positive or negative real number, which provide a number of benefits.
- The process of querying pairwise similarities is separated from the clustering algorithm (unlike previous work where the querying process is tightly integrated into the clustering algorithm itself).
- We propose an efficient local search algorithm for computing the correlation clustering solution (step 2 of the active learning procedure) which dynamically determines the number of clusters. This algorithm is robust to noise in the similarities, which is important in our setting since we have partial information about the ground-truth similarities.
- We assume a noisy oracle, where the noise is non-persistent, meaning that querying the same similarity more than once in order to correct for previous mistakes is possible. We propose two query strategies called **maxmin** and **maxexp** that take advantage of this (step 3 of the active learning procedure).

Results

The plot below shows the adjusted rand index (ARI) between the current clustering (step 2 of the AL procedure) and the ground-truth clustering. Maxexp and maxmin outperform all baseline methods. See [2] for details about all query strategies.



References

1. García-Soriano et al, Query-efficient correlation clustering (**WWW**) 2020.
2. Aronsson et al, Correlation Clustering with Active Learning of Pairwise Similarities TMLR 2024.

WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM