KU LEUVEN

DEPARTMENT OF COMPUTER SCIENCE

Tackling Noise in Active Semi-Supervised Clustering

Jonas Soenen[1], Sebastijan Dumančić[1], Toon Van Craenendonck[2] and Hendrik Blockeel[1]
[1]KU Leuven, [2]VITO
✉ jonas.soenen@cs.kuleuven.be

Key Idea

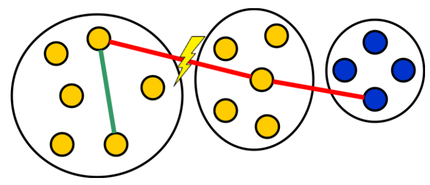
In active constraint-based clustering, some of **the user's answers might be noisy**
Reason probabilistically about the correctness of the user's answers
Ask additional questions to corroborate or correct suspicious answers

Background

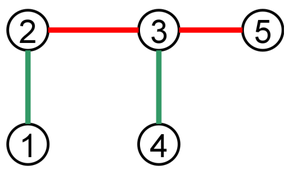
Active semi-supervised clustering exploits pairwise constraints to obtain clusterings that are better aligned with the user's interests. Instead of gathering constraints beforehand, actively ask the user for the constraint between an informative pair of instances.

Motivation

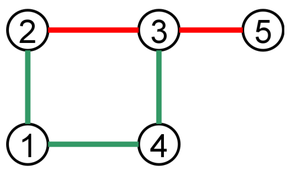
Many existing approaches **assume noiseless constraints**. However, in practice, the user might make **mistakes**. Forcing incorrect constraints to be satisfied can have **a detrimental impact** on the clustering quality.



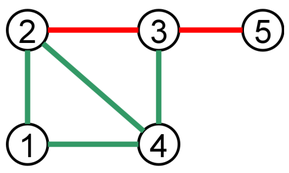
Approach



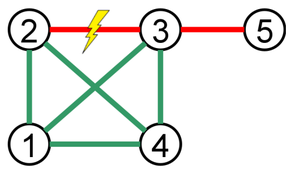
The confidence is only 0.81, ask redundant constraints



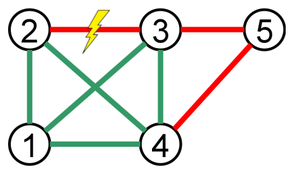
Inconsistent cycle, one of the involved constraints is noisy



The most likely noisy constraint is (2,3) or (1,4)



The most likely noisy constraint is (2,3)



The confidence is 0.90

Calculating the Confidence

We employ a **Bayesian approach** to calculate the probability that the ground truth constraints have values R given a set of constraints obtained from the user U:

$$P(\mathbf{R}\text{eal constraints}|\mathbf{U}\text{ser constraints}) = \frac{P(U|R)P(R)}{\sum_{R'} P(U|R')P(R')}$$

Assumptions:

- $P(R)$ is uniformly distributed over all consistent values of R (i.e. the constraint don't contradict each other)
- There is a fixed probability that the user answers a constraint incorrectly

Exact calculation infeasible ⇒ **approximate** by only considering the **most likely** consistent values of R

Selecting Informative Redundant Constraints

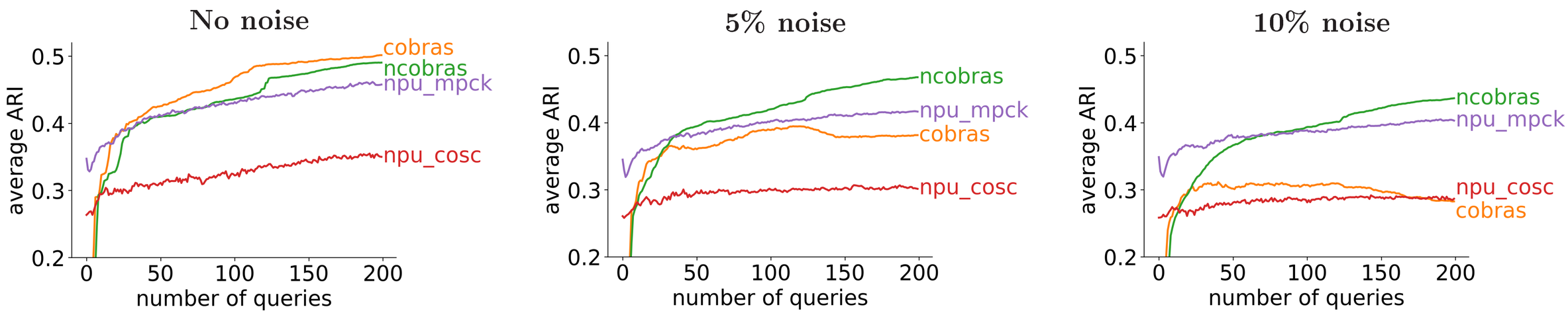
Corroborate If there is one most likely assignment R, query a constraint that will probably agree with R and disagree with other likely assignments

Correct If there are multiple most likely assignments, query a constraint that will probably agree with half of the most likely assignments and disagree with the other half

Noise-Robust COBRAS

COBRAS is a practical active constrained-based clustering algorithm. However, due to its low bias it is **very sensitive to noise**. Too illustrate the effectiveness of our method we use it to devise a **noise-robust** version of COBRAS (named nCOBRAS).

Experiments



nCOBRAS is noise robust nCOBRAS clustering quality is relatively stable over all amounts of noise

nCOBRAS is query-efficient nCOBRAS outperforms COBRAS in the presence of noise and both NPU-MPCKmeans and NPU-COSC for all levels of noise

KU Leuven
Department of Computer Science
Celestijnenlaan 200A box 2402
3001 Leuven, Belgium

This research received funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme.