Class Prior Estimation in Active Positive and Unlabeled Learning

KU LEUVEN · LEUVEN.AI INSTITUTE · DTAI Declarative Languages & Artificial Intelligence

Lorenzo Perini, Vincent Vercruyssen and Jesse Davis
name.surname@kuleuven.be

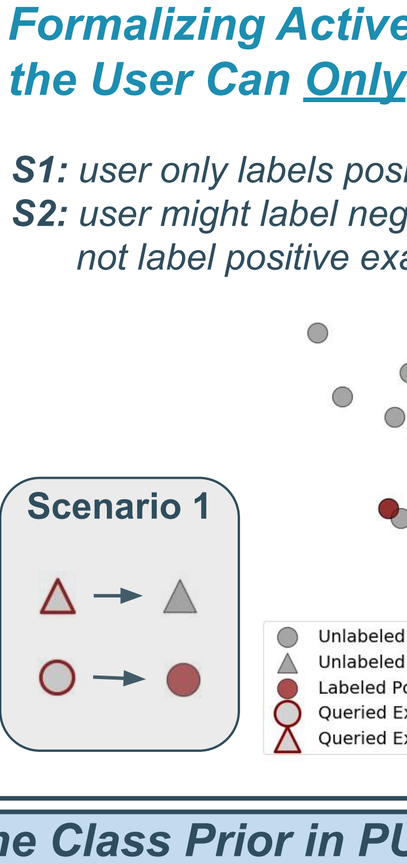
@LorenzoPerini95
https://github.com/Lorenzo-Perini/Active_PU_Learning

Positive and Unlabeled (PU) Data Arises Naturally. For Example:

***Task**: Estimate the relevance of ads based on click data*
***Challenge**: Ads the user has not clicked on may still be relevant!*

Positive (click)
Unlabeled (no click)

Common Approach to Learning from PU Data Revolves around Estimating the Class Prior



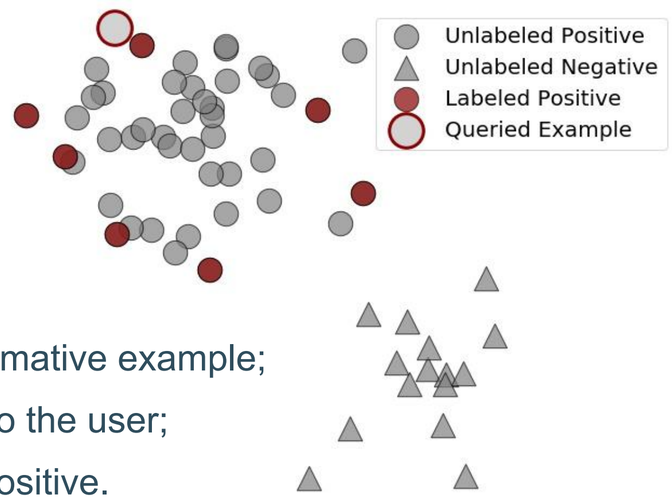
Most approaches estimate the class prior by assuming labels are Selected Completely At Random (**SCAR**)

$$\mathbb{P}(labeled) = c \cdot \mathbb{P}(positive)$$

c: label frequency

Contribution 1: Formalizing the Problem of PU Active Learning

Problem Setting: Acquiring Labels via Active Learning Violates SCAR Assumption



Active Learning Loop:

1. Train a model;
2. Select the most informative example;
3. Query the example to the user;
4. Collect the label, if positive.

SCAR assumption no longer holds as labels are selected via active learning, which introduces some bias

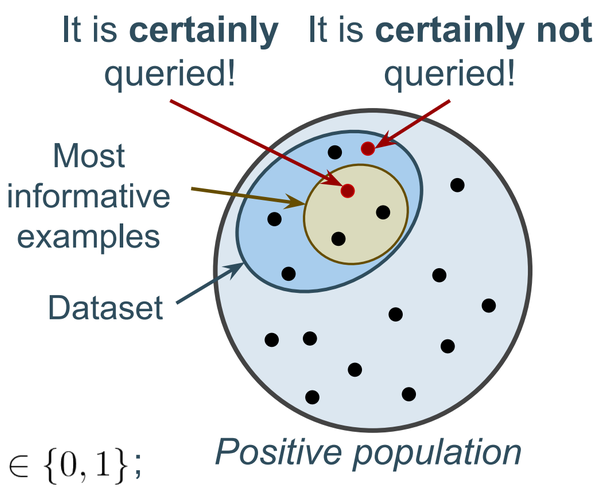
Formalizing Active Learning in a PU Setting: the User Can <u>Only</u> Provide Positive Labels

***S1:** user only labels positive examples.*
***S2:** user might label negative examples as positive or not label positive examples.*

Contribution 2: CAPe Is a Method For Estimating the Class Prior in PU Active Learning

When Collecting Labels Via Active Learning, Examples Contribute by their Propensity Score

Propensity Score: Instance specific probability to be labeled, if positive.



- **Given a dataset**, propensity score $e(x) \in \{0, 1\}$;
- Draw subsamples, apply combinatorics and average rules.

Deriving the Class Prior As Weighted Average of Positive Examples' Contribution

Labeled examples contributes fully

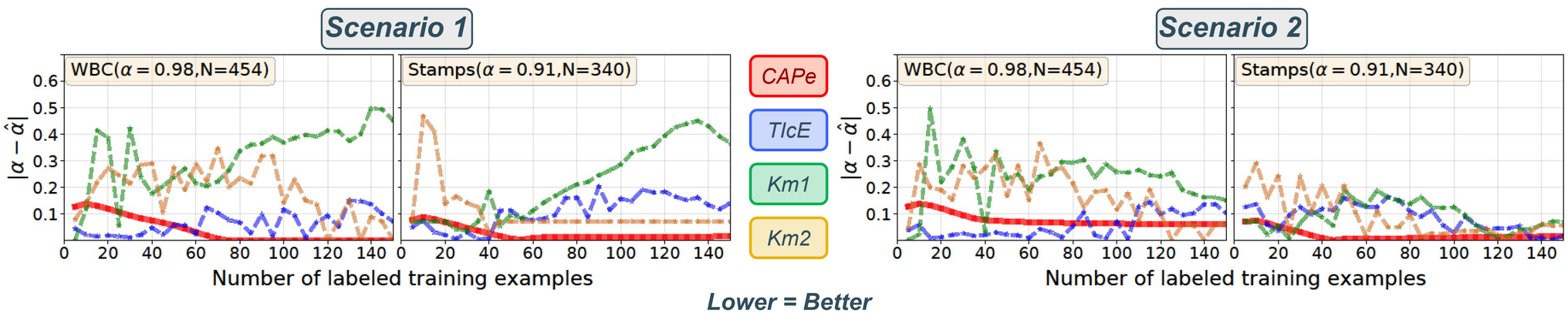
Unlabeled examples contribute their probability to be positive weighted by their propensity score

$$\mathbb{P}(positive) = \mathbb{E}_x\left[labeled\right] + \mathbb{E}_x\left[unlabeled \times \frac{\hat{y}(1 - e(x))}{1 - \hat{y}\, e(x)}\right]$$

- e(x) is the propensity score;
- $\hat{y} = \mathbb{P}\left(positive \,|\, x,\, e(x),\, h(x)\right) \in (0, 1)$;
- h is a learning model.

Contribution 3: Theoretical Results Show that CAPe Is Unbiased to the Limit

Contribution 4: CAPe Recovers the Class Prior Accurately in the Context of Anomaly Detection



Lower = Better