# Conference ML in PL 2022

## How to learn classifier chains using positive-unlabelled multi-label data?

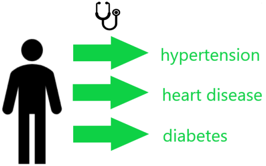
**Paweł Teisseyre**

- Institute of Computer Science, Polish Academy of Sciences
- Faculty of Mathematics and Information Sciences, Warsaw University of Technology

# Overview

1. Positive Unlabelled (PU) multi-label data
2. Modifications of classifier chains
3. Experiments

# Multi-label data

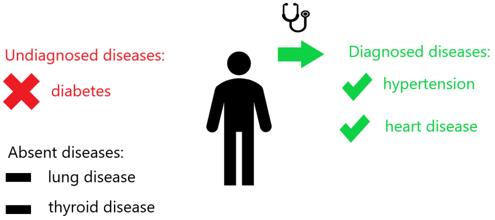



Absent diseases: lung disease, thyroid disease

- Labels: (hypertension, heart disease, diabetes, lung disease, thyroid disease)
- Label vector: $Y = (1, 1, 1, 0, 0)$
- Feature vector $X = (X_1, \ldots, X_p)$ (e.g. sex, age, diagnostic tests, genetic data, etc.)

Main goal: predict $Y$ using $X$.

# PU multi-label data




- Labels: (hypertension, heart disease, diabetes, lung disease, thyroid disease)
- True label vector: $Y = (1, 1, 1, 0, 0)$
- Observed label vector: $S = (1, 1, 0, 0, 0)$

# Positive unlabelled multi-label classification

- Vector of features $X = (X_1, \ldots, X_p)^T$.
- Vector of target variables (labels) $Y = (Y_1, \ldots, Y_K)^T$ **is not observed directly**.
- We observe $S = (S_1, \ldots, S_K)^T$ such that:
  - Value $S_k = 1$ means that $k$-th target is positive, i.e. $Y_k = 1$
  - Value $S_k = 0$ means that $k$-th target is not assigned ($Y_k = 1$ or $Y_k = 0$)
- Main goal: **build a model using training data which predicts $Y$ using $X$.**

# Positive unlabelled multi-label classification

- Training data consists of pairs $(x^{(i)}, s^{(i)})$ corresponding to $(X, S)$.
- We assume so-called **single data scenario**:
  - There is some unknown distribution $P(Y, X, S)$ such that $(x^{(i)}, y^{(i)}, s^{(i)})$, $i = 1, \ldots, n$ is i.i.d. sample drawn from it.
  - Only data $(x^{(i)}, s^{(i)})$ is observed.

## Positive unlabelled multi-label classification

**Important quantities**:

- **Label frequency** for $k$-th target variable: $c_k = P(S_k = 1 | Y_k = 1)$.
- Label frequency is related to class prior:

$$c_k = P(S_k = 1 | Y_k = 1) = \frac{P(S_k = 1, Y_k = 1)}{P(Y_k = 1)} = \frac{P(S_k = 1)}{P(Y_k = 1)}.$$

- It is easy to estimate $P(S_k = 1)$.
- Thus, it is easy to estimate accurately $c_k$, when class prior $\pi_k = P(Y_k = 1)$ is known.

# Classifier chains in multi-label classification [1] [2]

- Chain rule: $P(Y_1, \dots, Y_K|\mathrm{X}) = P(Y_1|\mathrm{X}) \prod_{k=2}^{K} P(Y_k|\mathrm{X}, Y_1, \dots, Y_{k-1})$.
- **Classifier chains (CC)**, chain of $K$ models:

$$
\begin{aligned}
&Y_1 \leftarrow X_1, \dots, X_p \\
&Y_2 \leftarrow X_1, \dots, X_p, Y_1 \\
&Y_3 \leftarrow X_1, \dots, X_p, Y_1, Y_2 \\
&\vdots \\
&Y_K \leftarrow X_1, \dots, X_p, Y_1, \dots, Y_{K-1}
\end{aligned}
$$

- **PROBLEM**: In the case of PU data, we do not observe $Y_1, \dots, Y_K$ directly.

[1] J. Read et. al., Classifier chains for multi-label classification, Machine Learning, 2011.

[2] J. Read et. al., Classifier Chains: A Review and Perspectives, J. Artif. Int. Res., 2020.

# **Method 1**: Naive classifier chains

- **Classifier chains (CC)**, chain of $K$ models:

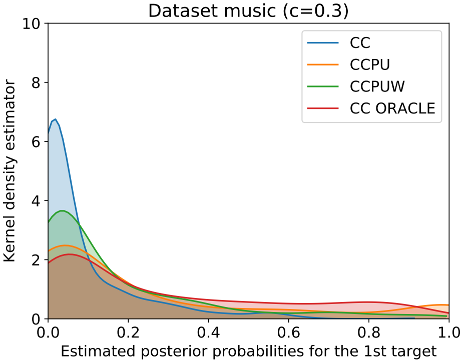
$$
\begin{aligned}
&S_1 \leftarrow X_1, \dots, X_p \\
&S_2 \leftarrow X_1, \dots, X_p, S_1 \\
&S_3 \leftarrow X_1, \dots, X_p, S_1, S_2 \\
&\vdots \\
&S_K \leftarrow X_1, \dots, X_p, S_1, \dots, S_{K-1}
\end{aligned}
$$

**PROBLEM**:

- We do not approximate conditional probabilities corresponding to the true target variables $P(S_k = 1|\mathrm{X}, S_1, \dots, S_{k-1}) \neq P(Y_k = 1|\mathrm{X}, Y_1, \dots, Y_{k-1})$.
- In particular, we have:

$$P(S_1 = 1|X) = \underbrace{P(S_1 = 1|X, Y_1 = 1)}_{\leq 1} P(Y_1 = 1|X) \leq P(Y_1 = 1|X).$$

# Estimated posterior probabilities




Rysunek: Smoothed histograms of estimated posterior probabilities for the first target in the chain, for $c_1 = P(S_1 = 1 | Y_1 = 1) = 0.3$.

In naive method (CC), estimated posterior probabilities are shrinked towards 0.

# PU Multi-label classification

## Selected Completely at Random (SCAR) assumption

For each $k = 1, \ldots, K$

$$P(S_k = 1 | \mathrm{X}, Y_k = 1, \mathrm{Y}_{A_{-k}}) = P(S_k = 1 | Y_k = 1),$$

for any subset $A_{-k} \subset \{1, \ldots, K\} \setminus \{k\}$.

## Fact

Under SCAR assumption we have, for any subset $A_{-k} \subset \{1, \ldots, K\} \setminus \{k\}$

$$P(Y_k = 1 | \mathrm{X}, \mathrm{Y}_{A_{-k}}) = c_k^{-1} P(S_k = 1 | \mathrm{X}, \mathrm{Y}_{A_{-k}}).$$

## **Method 2**: Classifier chains for PU data (CCPU)

Input: $X, S$, prior probabilities $\pi_1, \ldots, \pi_k$.

1. Estimate $c_k$ using equation $c_k = P(S_k = 1)/\pi_k$.
2. In $k$-th step:
   1. Fit model $S_k \leftarrow \mathrm{X}, \hat{Y}_1, \ldots, \hat{Y}_{k-1}$ to estimate $P(S_k = 1|\mathrm{X}, Y_1, \ldots, Y_{k-1})$.
   2. Estimate $P(Y_k = 1|\mathrm{X}, Y_1, \ldots, Y_{k-1})$ using equation $P(Y_k = 1|\mathrm{X}, Y_1, \ldots, Y_{k-1}) = c_k^{-1} P(S_k = 1|\mathrm{X}, Y_1, \ldots, Y_{k-1})$.
   3. Make prediction of $Y_k$, denoted as $\hat{Y}_k$, using estimate of $P(Y_k = 1|\mathrm{X}, Y_1, \ldots, Y_{k-1})$.

# **Method 3**: Classifier chains for PU data (CCPUW)

The risk associated with $k$-th classifier $g_k$ in the chain:

$$R(g_k) = E_{Z_k, Y_k} L(g_k(Z_k), Y_k) = \\ \alpha_k E_{Z_k | Y_k = 1} L^+(g_k(Z_k)) + (1 - \alpha_k) E_{Z_k | Y_k = 0} L^-(g_k(Z_k)).$$

where:

- $Z_k = (X, Y_1, \ldots, Y_{k-1})$
- $\alpha_k = P(Y_k = 1)$ (class prior for $k$-th target)
- $L^+$ and $L^-$ are are losses for positive and negative examples, respectively.

# **Method 3**: Classifier chains for PU data (CCPUW)

**Theorem**

*Let* $\alpha_k = P(Y_k = 1)$ *and* $c_k = P(S_k = 1|Y_k = 1)$ *be the label frequency for k-th label. The following equality holds*

$$R(g_k) = c_k\alpha_k E_{Z_k|S_k=1}\left[\frac{1}{c_k}L^+(g_k(Z_k)) + (1 - \frac{1}{c_k})L^-(g_k(Z_k))\right]$$
$$+(1 - c_k\alpha_k)E_{Z_k|S_k=0}L^-(g_k(Z_k))$$

- The optimal classifier for $k$-th target is defined as $g_k^* := \arg\min_{g_k} \hat{R}(g_k)$.

## **Method 3**: Classifier chains for PU data (CCPUW)

Input: $X, S$, prior probabilities $\pi_1, \ldots, \pi_k$.

1. Estimate $c_k$ using equation $c_k = P(S_k = 1)/\pi_k$.
2. First step:
   1. Fit classifier $g_1^*$ to estimate $P(Y_k = 1|\mathrm{X})$.
   2. Make prediction of $Y_1$, denoted as $\hat{Y}_1$.
3. In the $k$-th step:
   1. Fit classifier $g_k^*$ using X, $\hat{Y}_1, \ldots, \hat{Y}_{k-1}$ as features to estimate $P(Y_k = 1|\mathrm{X}, Y_1, \ldots, Y_{k-1})$.
   2. Make prediction of $Y_k$, denoted as $\hat{Y}_k$.

# Experiments

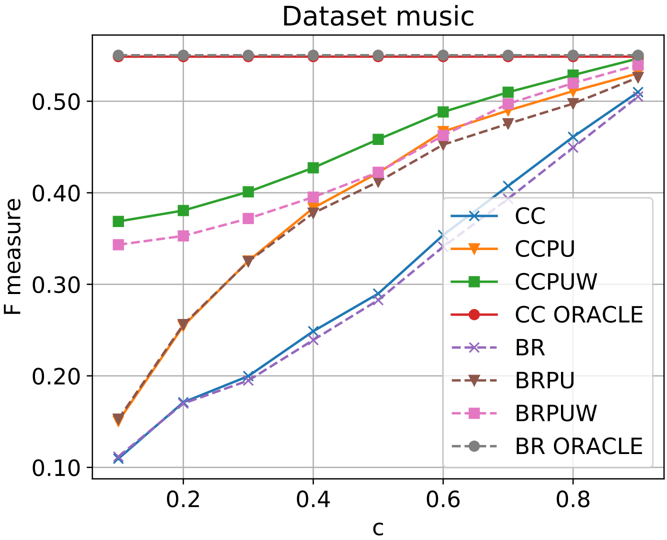
Datasets:

1. We created PU datasets from the original completely labelled datasets in the following way.
2. For each target variable, the positive examples (wrt to this target) are selected to be labelled with label frequency $c$, where $c$ is treated as a parameter which varies in the experiments.

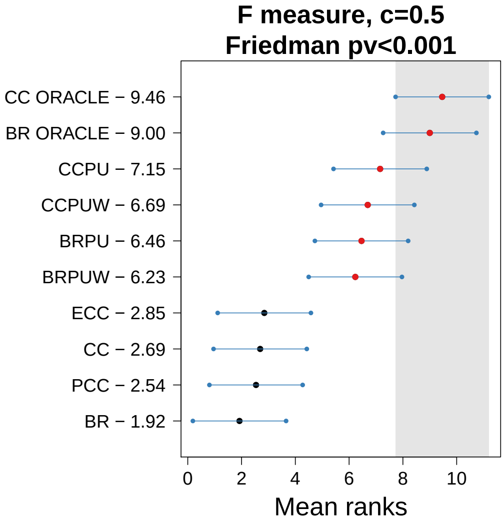
Methods:

1. Oracle method: CC ORACLE
2. Naive methods: CC
3. Proposed methods: CCPU, CCPUW.
4. Corresponding Binary Relevance (BR) methods: BR ORACLE, BR, BRPU, BRPUW

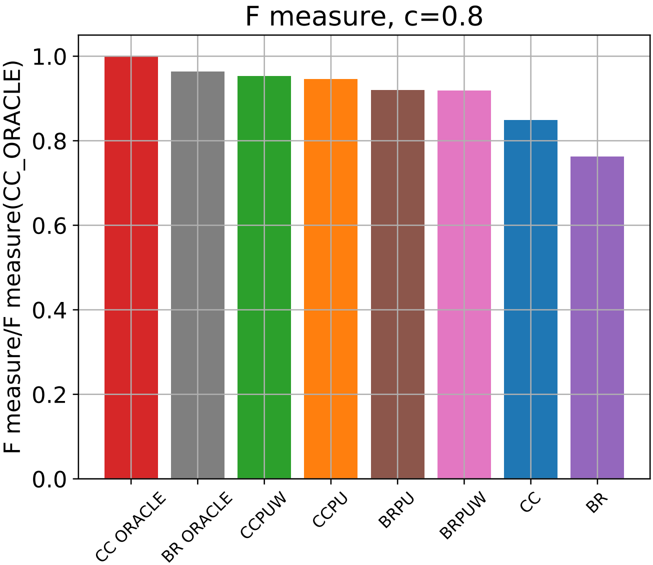
# How prediction accuracy depends on $c$?



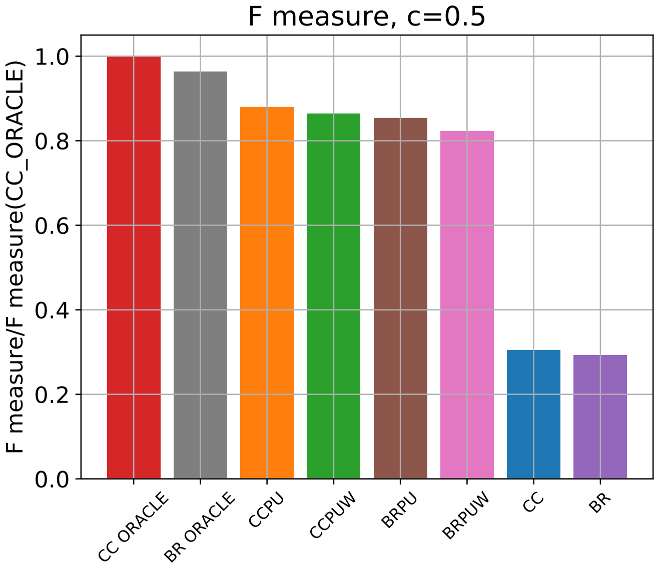
# Results of Friedman and pairwise tests



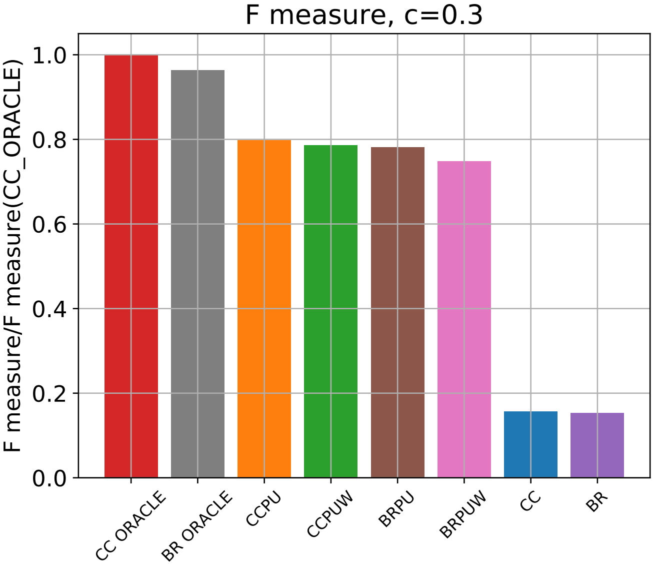
## How much we lose compared to the optimal method?



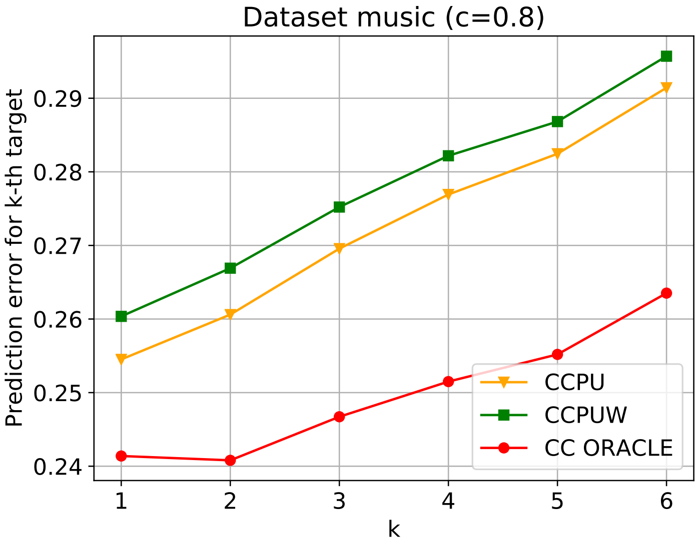
How much we lose compared to the optimal method?



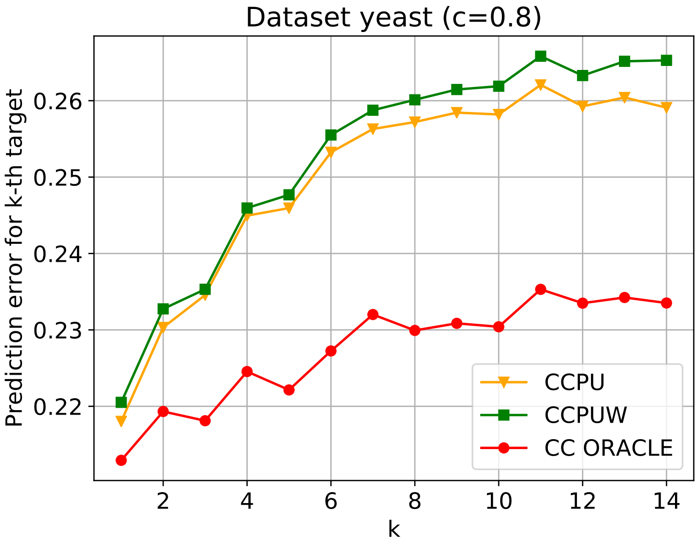
How much we lose compared to the optimal method?



# Error propagation in classifier chains



# Error propagation in classifier chains



# Conclusions

- PU multi-label problem is challenging (dependencies between observed target variables may be much weaker than between original ones).
- Building classifier chains for PU multi-label is also challenging (noisy target variables and noisy features).
- Naive method works poorly.
- The performance of the considered method deteriorates for small label frequency.
- The proposed methods work significantly better than naive method, although they are still worse than ORACLE methods, especially for small $c$.
- The differences between CC-based methods and BR-based methods are not very pronounced.

# References


1. P. Teisseyre, *Classifier chains for positive unlabelled multi-label learning*, Knowledge-Based Systems, 2021.
2. J. Read et. al., *Classifier chains for multi-label classification*, Machine Learning, 2011.
3. J. Read et. al., Classifier Chains: A Review and Perspectives, J. Artif. Int. Res., 2020.
4. J. Bekker, J. Davis, *Learning from positive and unlabeled data: a survey*, Machine Learning, 2020.
5. P. Teisseyre, J. Mielniczuk, M. Łazęcka, *Different strategies of fitting logistic regression for positive unlabeled data*, Proceedings of ICCS'20, 2020.