# One-class classification approach to variational learning from biased Positive Unlabeled data

**Jan Mielniczuk**

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

Based on joint research with **A. Wawrzeńczyk**

# Outline

1. Introduction: biased Positive Unlabeled data
2. Autoencoders
   VaDE (Jiang et al 2017)
   VAE-PU (Na et al. 2020)
3. Our contribution: extension of VAE-PU: VAE-PU +OCC

# PU datasets: partial observability in action

Table: Texting while driving survey - obtained data

| Age | Gender | Education | Survey answer | Texts |
|---|---|---|---|---|
| 20 | male | higher | no | ? |
| 50 | female | primary | yes | yes |
| 35 | female | secondary | no | ? |
| 15 | male | primary | no | ? |
| 70 | male | secondary | no | ? |
| 30 | female | primary | yes | yes |

Many examples in medicine, biology, NLP (text annotation) etc.

# PU learning task

Instead of $(X, Y)$ $(Y = 1, -1$, positve, negative) we observe $(X, O)$ $(O = 1, 0$ (labeled, unlabeled)

**Positive-Unlabeled (PU) learning**:

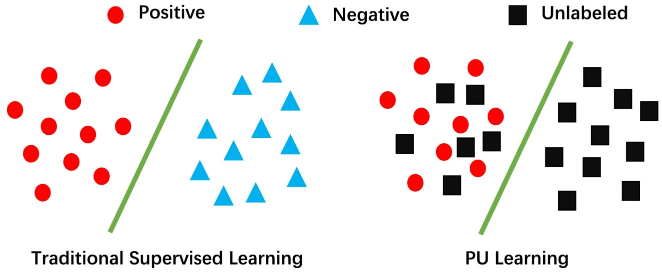
- Labeled and unlabeled sample ($O$ - label vector),
- All **labeled** observations are **positive**,
- **Unlabeled** observations can be **positive or negative**.

We want to to build a classifier $\hat{Y}$ of true class indicator $Y$ and estimate posterior probability

$$y(x) := P(Y = 1|x)$$

# Positive and unlabelled data

Visualization of traditional classification and classification from PU data [1]




[1] Gong et, al., IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019.

# Labeling assumptions

**Propensity score**:

$$e(x) := P(O = 1 | Y = 1, x)$$

**Selected Completely At Random (SCAR)** assumption:

$$e(x) = P(O = 1 | Y = 1, {\color{red}x}) = P(O = 1 | Y = 1) = const.$$

$c = P(O = 1 | Y = 1)$ is the **label frequency**.

**Selected At Random (SAR)** assumption:

$$e(x) = P(O = 1 | Y = 1, {\color{red}x})$$

Weaker **SAR** (Selected At Random) assumption can be used instead:

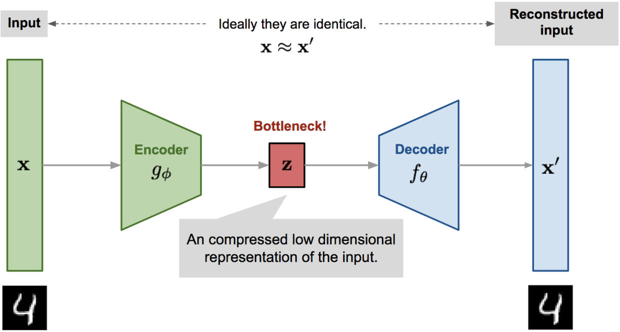
$$e(x) = P(O = 1|Y = 1, x)$$

Propensity score is a function of object attributes (biased PU data)!

Current advances in biased PU modeling:

- EM Bekker, Davis (2017),
- **VAE-PU** Na et al (2020),
- LBE Gong et al (2021),
- JOINT, TWO MODELS Furmańczyk, JM, Rejchel, Teisseyre (2021),

# Autoencoders , Variational Auto-Encoders – idea

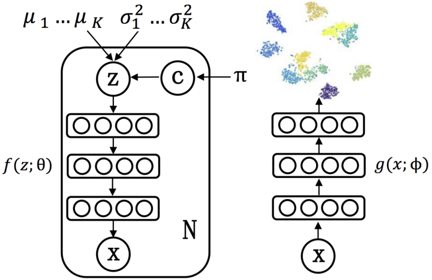



Latent space of traditional autoencoders is not regularised.
Variational Auto-Encoders: introduction of variational distribution $q(z, x)$ and maximisation of Evidence Lower BOund (ELBO).

# VaDE: idea

**Variational Deep Embedding (VaDE)** (Jiang et al (2017)).

**Idea:** Model latent variable $z$ as the mixture of gaussians.



**Generative process:**

- Choose a cluster $c \sim \text{Cat}(\pi)$
- Choose a latent vector $z \sim \mathcal{N}(\mu_c, \sigma_c^2 I)$
- Generation of $x$ (real number case):
  - Compute $\mu_x$ and $\sigma_x^2$
    $$[\mu_x, \log \sigma_x^2] = f(z; \theta)$$
  - Choose an observation $x \sim \mathcal{N}(\mu_x, \sigma_x^2 I)$

**Generative process:**

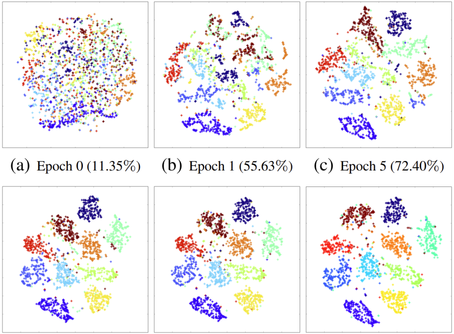
- Choose a cluster $c \sim \mathrm{Cat}(\pi)$
- Choose a latent vector $z \sim \mathcal{N}(\mu_c, \sigma_c^2 I)$
- Generation of $x$ (real number case):
  - Compute $\mu_x$ and $\sigma_x^2$

$$[\mu_x, \log \sigma_x^2] = f(z; \theta)$$

  - Choose an observation $x \sim \mathcal{N}(\mu_x, \sigma_x^2 I)$

Variational posterior $q(z, c|x) = q(z|x)q(c|x) \Rightarrow$ ELBO bound.

# Results for MNIST



(a) Epoch 0 (11.35%) (b) Epoch 1 (55.63%) (c) Epoch 5 (72.40%)

(d) Epoch 50 (84.59%) (e) Epoch 120 (90.76%) (f) Epoch End (94.46%)

Figure: Colors: ground truth classes,clusters are given by latent encoding, t-SNE representation [2]

[2] Jiang et al., IJCAI'2017

# VAE-PU: empirical risk minimisation

$y \in \{-1, 1\}$, $g(x)$ - target classifying function (e.g. neural network), $l(\cdot)$ - any loss function (eg. sigmoid), $o$ - label vector.
A **general PU risk function**:

$$R_{PU}(g) = p(y = +1, o = 1)\mathbb{E}_{x \sim p_{pl}(x)}[l(g(x))]$$
$$+ p(y = +1, o = 0)\mathbb{E}_{x \sim p_{pu}(x)}[l(g(x)) - l(-g(x))]$$
$$+ p(o = 0)\mathbb{E}_{x \sim p_{u}(x)}[l(-g(x))]$$

# VAE-PU: empirical risk minimisation

$y \in \{-1, 1\}$, $g(x)$ - target classifying function (e.g. neural network), $l(\cdot)$ - any loss function (eg. sigmoid), $o$ - label vector.
A **general PU risk function**:

$$R_{PU}(g) = p(y = +1, o = 1)\mathbb{E}_{x \sim p_{pl}(x)}[l(g(x))] + p(y = +1, o = 0)\mathbb{E}_{x \sim p_{pu}(x)}[l(g(x)) - l(-g(x))] + p(o = 0)\mathbb{E}_{x \sim p_u(x)}[l(-g(x))]$$

**Notation**

$$\pi = P(Y = 1) \text{ assumed known}$$

$$\pi_{PL} = P(Y = 1, O = 1),\ \pi_{PU} = P(Y = 1, O = 0)$$

# Empirical risk

Empirical risk function:

$$\hat{R}_{PU}(g) = \frac{\pi_{PL}}{|\chi_{PL}|} \sum_{x^{(pl)} \in \chi_{PL}} l(g(x^{(pl)}))$$

$$+ \frac{\pi_{PU}}{|\tilde{\chi}_{PU}|} \sum_{\tilde{x}^{(pu)} \in \tilde{\chi}_{PU}} l(g(\tilde{x}^{(pu)}))$$

$$+ \max\left\{0, -\frac{\pi_{PU}}{|\tilde{\chi}_{PU}|} \sum_{\tilde{x}^{(pu)} \in \tilde{\chi}_{PU}} l(-g(\tilde{x}^{(pu)})) + \frac{\pi_U}{|\chi_U|} \sum_{x^{(u)} \in \chi_U} l(-g(x^{(u)}))\right\}$$

**Problem:** We need to estimate the distribution of PU cases (due to terms with $\tilde{\chi}_{PU}$).
**Idea:** Use model similar to VaDE to generate PU pseudo-observations.

# Generative process

Instead of one latent representation $z$, we use **two latent vectors**:

- $h_o$ - encodes **observation** status (labeled,unlabeled),
- $h_y$ - encodes **class** information (positive, negative) .

Motivation: positive cases, regardless of what is observed, share the same $h_y$.

**Generative process**:

- Choose cluster $c \sim \text{Bern}(\eta)$
- Generate latent class vector $h_y | c \sim \mathcal{N}(\mu_c, \sigma_c^2 I)$
- Generate latent observation vector $h_o \sim \mathcal{N}(0, I)$
- Generate sample $x$:
  - $[\mu_x, \log \sigma_x^2] = f(h_y, h_o; \theta)$
  - $x | h_y, h_o \sim \mathcal{N}(\mu_x, \sigma_x^2 I)$
- Generate observation status $o | h_o \sim \text{Bern}(f_o(h_o))$

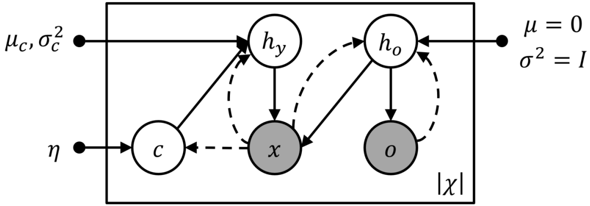



**Figure 2: The graphical model of the VAE-PU. The solid lines denote the generative model $p$ and the dashed lines denote the variational approximation $q$ to $p$. The gray and white circles denote the observed variables and latent variables, respectively. $|\chi|$ is the number of entire data instances.**

Joint probability can be factorized:

$$p(h_y, h_o, c, x, o) = p(c)p(h_y|c)p(h_o)p(o|h_o)p(x|h_y, h_o)$$

$$q(h_y, h_o, c|x, o) = q(h_y|x)q(h_o|x, o)q(c|x) \Rightarrow \text{ELBO bound}$$

# $h_y$ for Odd v Even (MNIST)

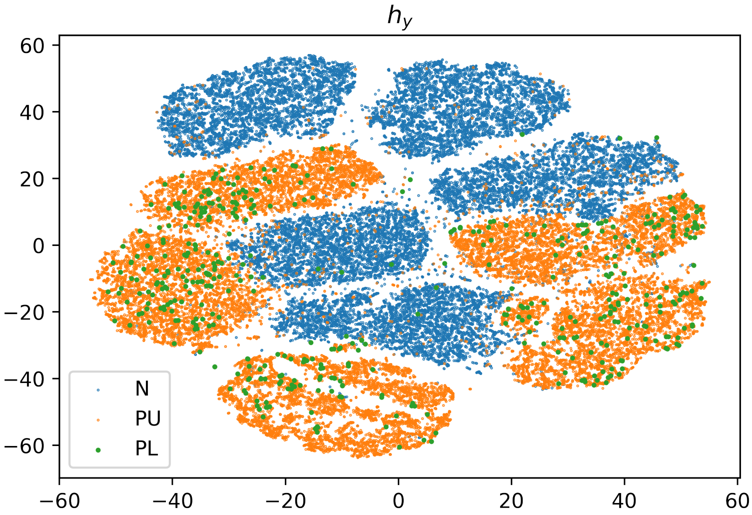



Figure: $h_y$ latent space, OvE, t-SNE representation

# Generation of artificial PU examples - crucial point !

In order to **generate PU pseudo-examples**:

1. Match positive and unlabeled samples (eg. nearest $h_y$ representation),
2. Extract label information from positive instance ($h_y^{(pl)}$) and observation status from unlabeled sample ($h_o^{(u)}$),
3. Concatenate $h_y^{(pl)}$ and $h_o^{(u)}$,
4. Decode the latent representation.
5. Constructed examples *mimic* elements of $\chi_{PU}$

# Generated examples

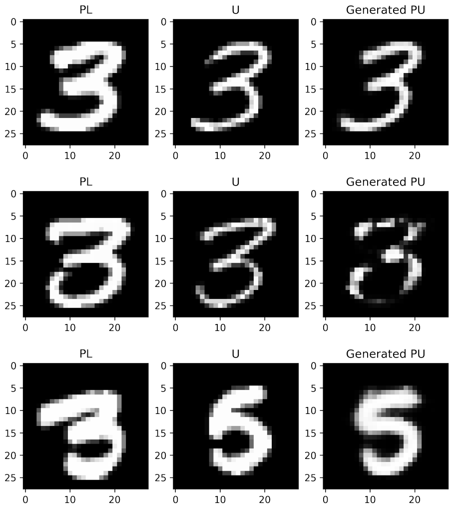



Table: Mean digit boldness

| Dataset | Boldness |
|---|---|
| PL | 0.2475 |
| True PU | **0.1397** |
| U | 0.1346 |
| **Generated PU** | **0.1451** |

# One-class classification

Idea: instead of using artificially constructed $\tilde{\chi}_{PU}$ in minimisation of empirical risk we try to extract PU examples *from U* using **one-class classification** methods. Having $\hat{\chi}_{PU} \subset \chi_U$ is advantageous.

# One-class classification

Idea: instead of using artificially constructed $\tilde{\chi}_{PU}$ in minimisation of empirical risk we try to extract PU examples *from U* using **one-class classification** methods. Having $\hat{\chi}_{PU} \subset \chi_U$ is advantageous.

**One-class classification (OCC, aka ODD aka Anomaly Detection)**:

- Training dataset $\mathcal{D} = \{X_i\}_{i=1}^{n}$ – iid. observations from unknown distribution $P_X$ (samples drawn from $P_X$ are **inliers**),
- Goal: test which among new set $\mathcal{D}^{test} = \{X_{n+i}\}_{i=1}^{n_{test}}$ are **outliers**, that is they are not drawn from the same distribution $P_X$.

# One-class classification

Idea: instead of using artificially constructed $\tilde{\chi}_{PU}$ in minimisation of empirical risk we try to extract PU examples *from U* using **one-class classification** methods. Having $\hat{\chi}_{PU} \subset \chi_U$ is advantageous.

**One-class classification (OCC, aka ODD aka Anomaly Detection)**:

- Training dataset $\mathcal{D} = \{X_i\}_{i=1}^{n}$ – iid. observations from unknown distribution $P_X$ (samples drawn from $P_X$ are **inliers**),
- Goal: test which among new set $\mathcal{D}^{test} = \{X_{n+i}\}_{i=1}^{n_{test}}$ are **outliers**, that is they are not drawn from the same distribution $P_X$.

Multiple known methods, eg.:

- One-Class SVM (Schölkopf et al 2001,
- Isolation Forest (Li et al.2008),
- ECOD (Liu et al. 2022)
- $A^3$: Activation Anomaly Analysis, Sperl et al. 2021)

# Algorithm VAE-PU +OCC (simplified)

Application in our setting: $\tilde{\chi}_{PU}$ are treated as inliers, outliers $\chi_{NU} \subseteq \chi_U$.

# Algorithm VAE-PU +OCC (simplified)

Application in our setting: $\tilde{\chi}_{PU}$ are treated as inliers, outliers $\chi_{NU} \subseteq \chi_U$.

**Algorithm**[a]

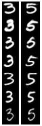
- Given classifying function $g$ train VAE-PU model optimise objective function to obtain pseudo-sample $\tilde{\chi}_{PU}$;
- Given $\tilde{\chi}_{PU}$ perform OCC to extract inliers $\hat{\chi}_{PU} \subseteq \chi_U$;
- Perform minimisation of empirical risk $R(g)$ with $\hat{\chi}_{PU}$ replacing $\tilde{\chi}_{PU}$;
- Perform the next cycle until $F1$ measure levels off.

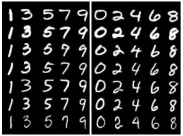
[a] https://github.com/adamw00000/VAE-PU-OCC

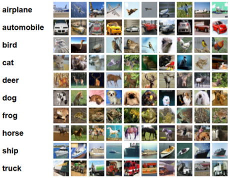
# Experimental settings

Datasets:

- MNIST: 3v5, OvE,
- CIFAR: CarTruck, MachineAnimal,
- STL (MachineAnimal),
- Gas concentrations.








Alternative methods:

- **Baseline: VAE-PU** (Na et al. 2020),
- **SAR-EM** (Bekker, Davis 2019),
- **LBE** (Gong et al., 2021).

Comparisons in the original paper:

- nnPU,
- uPU,
- PUbN/N,
- GenPU,
- PAN,
- PUSB.

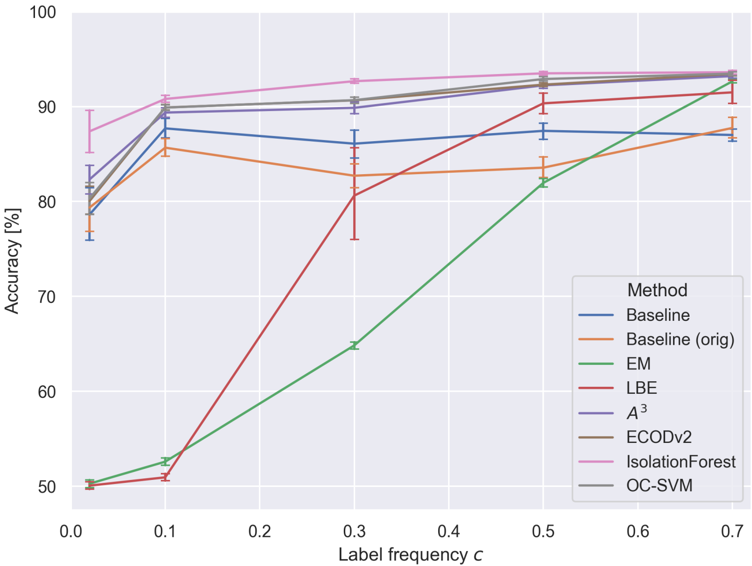
# Details of numerical experiments

- MNIST: two different tasks: 3 versus 5 (3v5) and Odds versus Evens (OvE)
  CIFAR-10, STL-10: Machine versus Animal
  Gas Concentrations: Ethanol versus Amonia

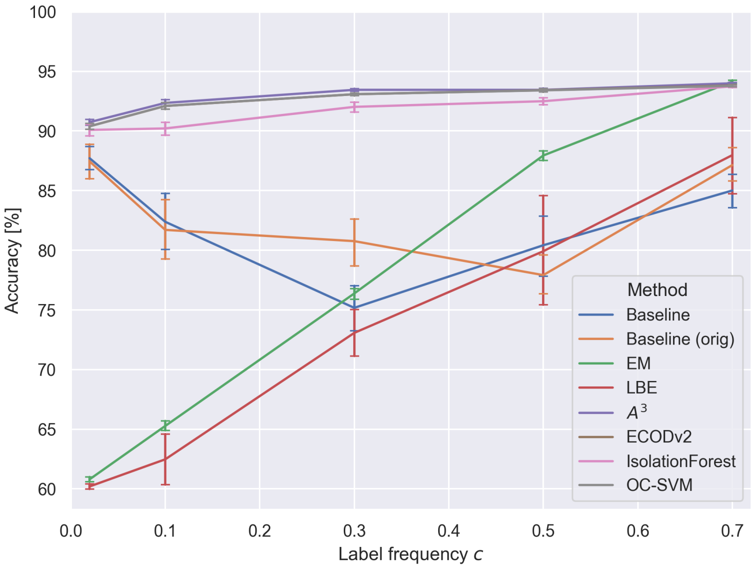
# Details of numerical experiments

- MNIST: two different tasks: 3 versus 5 (3v5) and Odds versus Evens (OvE)
  CIFAR-10, STL-10: Machine versus Animal
  Gas Concentrations: Ethanol versus Amonia
- Data labeled artificially according to various labeling scenarios:
  MNIST data: proportional to boldness, CIFAR-10, STL-10: proportional to 'redness'.
  Number of examples to be labeled is consistent with assumed label frequency $c = P(S = 1|Y = 1)$.

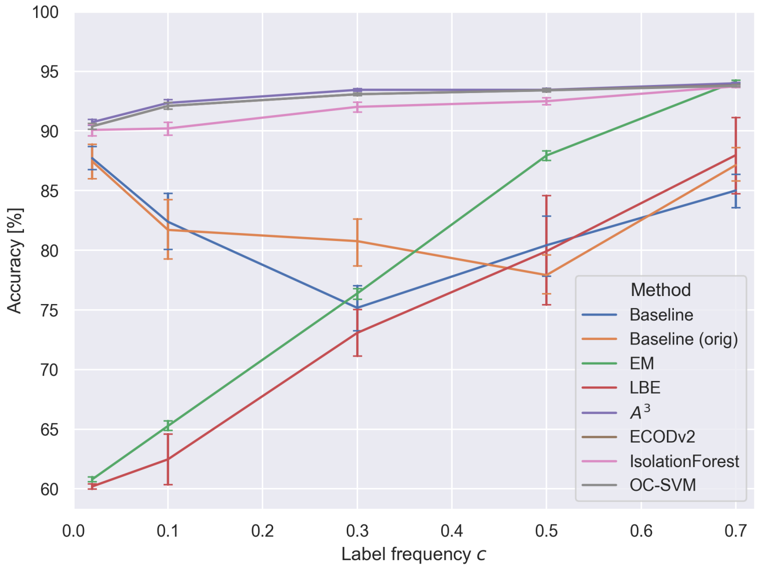
# Results:CIFAR CarTruck



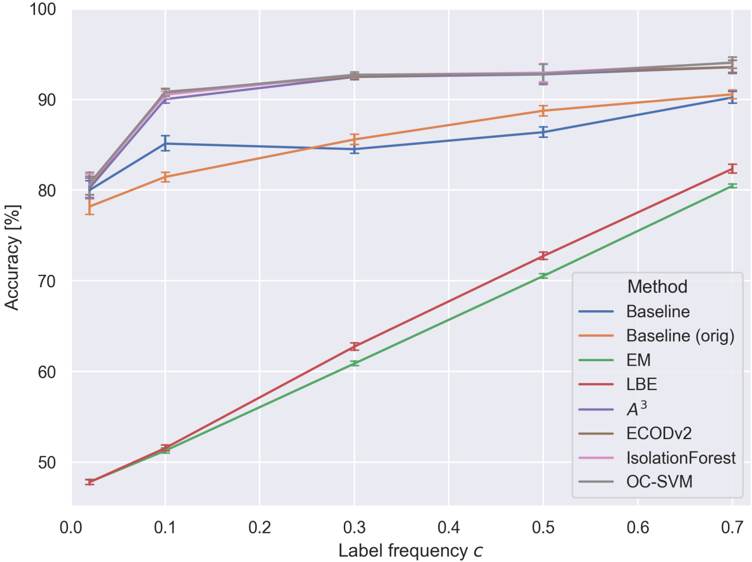
# Results:CIFAR MachineAnimal



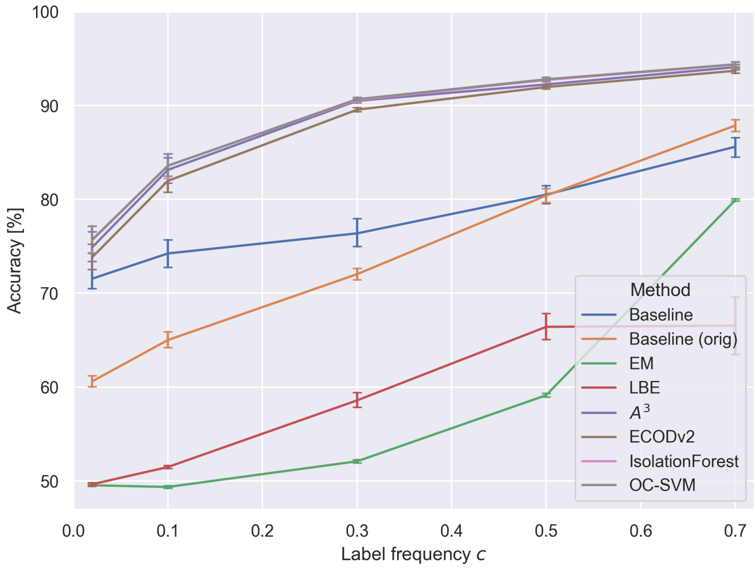
# Results:CIFAR MachineAnimal



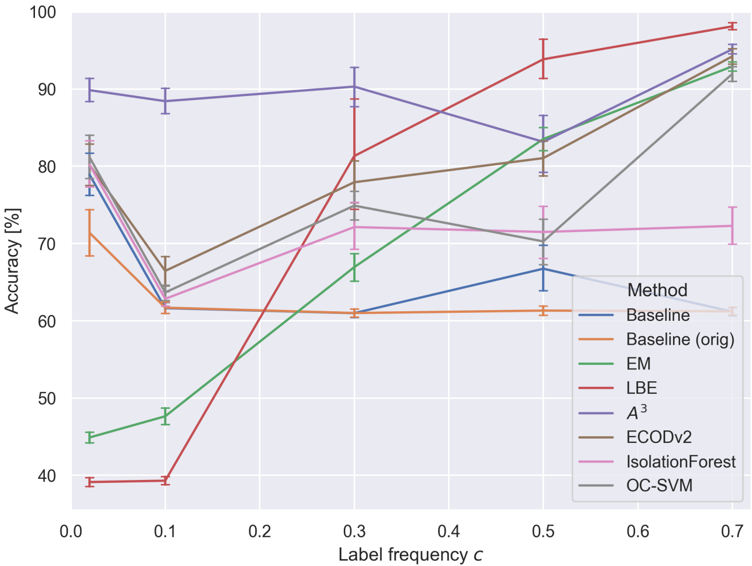
# Results:MNIST 3v5



# Results:MNIST OvE



# Results:Gas Concentrations



# Summary: VAE-PU+OCC

Conclusions from experiments:

- OCC modification **improved results significantly** as compared to baseline VAE-PU model,
- $A^3$ and *ECOD* variants perform **consistently the best** among OCC methods studied,
- EM and LBE methods **rarely** outperform OCC-enhanced model.
- EM and LBE methods work poorly for small labeling probability $c$.

# References


1. **LBE**:Gong, C. et al. Instance-Dependent Positive and Unlabeled Learning with Labeling Bias Estimation, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021
2. **VAE PU**: Na, B. et al., Deep Generative Positive-Unlabeled Learning under Selection Bias, CIKM 2020
3. **EM**: Bekker et al., Beyond the SCAR assumption for learning from positive and unlabeled data, ECML 2019
4. **VAE PU +OCC**: Wawrzeńczyk, A. and JM, One-class classification approach to variational learning from biased positive unlabeled data, 2022, submitted
5. **ECOD**: Li, Z. et al. ECOD: Unsupervised Outlier Detection Using Empirical Cumulative Distribution Functions, IEEE Transaction on Knowledge and Data Engineeering, 2022