Supplementary material to „One-Class Classification Approach to Variational Learning from Biased Positive Unlabeled Data”

Jan Mielniczuk[a,b;*] **and Adam Wawrzeńczyk**[a]

[a]Institute of Computer Science, Polish Academy of Sciences
[b]Faculty of Mathematics and Information Science, Warsaw University of Technology
ORCiD ID: Jan Mielniczuk https://orcid.org/0000-0003-2621-2303,
Adam Wawrzeńczyk https://orcid.org/0000-0002-6202-7829

A Matching of positive and unlabeled examples in VAE-PU

During sample generation process, positive and unlabeled examples have to be matched. Using values of latent variable h_y for labeled data, denoted by $h_y^{(pl)}$ we can construct pseudo-sample pertaining to P_{PU} in the following two ways:

(i) for each element of labeled sample with corresponding vector $\left(h_y^{(pl)}, h_s^{(pl)}\right)$, one finds an element among unlabeled examples with $\left(h_y^{(u)}, h_s^{u)}\right)$ such that $h_y^{(u)}$ is the closest to $h_y^{(pl)}$ wrt the euclidean distance (h_y-matching) and creates a latent vector $\left(h_y^{(pl)}, h_s^{u)}\right)$ which mimics latent (h_y, h_s) vector of a potential positive unlabeled observation. From this vector a pseudo-observation from positive unlabeled (PU) population is reconstructed by decoder part of VAE,

(ii) the above procedure is repeated, but now the matching element from unlabeled data is chosen by picking $\left(h_y^{(u)}, h_s^{u)}\right)$ such that $h_s^{(u)}$ is closest to $h_s^{(pl)}$ (h_s-matching). The remaining part of the construction is the same.

Table A1: Comparison of different latent representation matching algorithms, MNIST 3v5.

Dataset	Mean digit boldness
PL	0.2475
U	0.1346
True PU	**0.1397**
Generated PU – $\boldsymbol{h_s}$-matching	**0.2255**
Generated PU – $\boldsymbol{h_y}$-matching	**0.1451**

VAE-PU paper [16] does not assume any particular matching method, only indicating h_y matching (i) in its text as one of the possibilities. The matching algorithm used in its official implementation is the h_o matching (ii). We believe that generating PU samples in this

* Corresponding Author. Email: jan.mielniczuk@ipipan.waw.pl.

Table A2: Benchmark dataset statistics

Name	Samples	Features	Class prior π
MNIST 3v5	13454	784	0.53
MNIST OvE	70000	784	0.51
CIFAR CarTruck	12000	512	0.50
CIFAR MachineAnimal	60000	512	0.40
STL MachineAnimal	13000	512	0.40
Gas Concentrations	4206	129	0.61

Table A3: MNIST 3v5 results – no-SCAR.

c	Method	Accuracy	Precision	Recall	F1 score
0.02	Baseline	79.99 ± 1.04	77.27 ± 0.99	**88.75 ± 0.85**	**82.59 ± 0.85**
	Baseline (orig)	78.18 ± 0.97	75.65 ± 0.98	87.06 ± 1.17	80.91 ± 0.85
	LBE	47.78 ± 0.29	**94.95 ± 1.75**	1.86 ± 0.19	3.64 ± 0.37
	SAR-EM	47.79 ± 0.31	94.33 ± 1.63	1.51 ± 0.14	2.96 ± 0.28
	A^3	80.21 ± 1.07	78.20 ± 1.10	87.50 ± 0.82	82.56 ± 0.86
	ECODv2	80.44 ± 1.08	78.53 ± 1.23	87.18 ± 0.86	**82.59 ± 0.92**
	IsolationForest	80.63 ± 1.16	80.04 ± 1.45	85.37 ± 1.01	82.53 ± 0.92
	OC-SVM	**80.73 ± 1.23**	80.57 ± 1.68	84.86 ± 1.01	82.54 ± 0.96
0.10	Baseline	85.11 ± 0.87	79.35 ± 1.06	**97.61 ± 0.27**	87.50 ± 0.63
	Baseline (orig)	81.44 ± 0.57	76.80 ± 0.72	93.42 ± 0.84	84.26 ± 0.47
	LBE	51.54 ± 0.35	93.77 ± 0.63	9.47 ± 0.45	17.18 ± 0.74
	SAR-EM	51.25 ± 0.31	**94.45 ± 0.66**	8.82 ± 0.26	16.13 ± 0.43
	A^3	90.01 ± 0.47	90.45 ± 0.69	90.88 ± 0.41	90.65 ± 0.39
	ECODv2	**90.82 ± 0.37**	91.69 ± 0.61	91.06 ± 0.72	**91.34 ± 0.32**
	IsolationForest	90.52 ± 0.41	90.83 ± 0.54	91.44 ± 0.56	91.12 ± 0.36
	OC-SVM	90.75 ± 0.43	91.42 ± 0.62	91.23 ± 0.54	91.30 ± 0.37
0.30	Baseline	84.50 ± 0.50	77.99 ± 0.64	**98.83 ± 0.15**	87.16 ± 0.36
	Baseline (orig)	85.57 ± 0.59	80.66 ± 0.88	96.01 ± 0.30	87.64 ± 0.44
	LBE	62.72 ± 0.44	92.85 ± 0.49	32.39 ± 0.69	47.99 ± 0.75
	SAR-EM	60.85 ± 0.27	**95.09 ± 0.47**	27.81 ± 0.37	43.02 ± 0.43
	A^3	92.47 ± 0.32	93.85 ± 0.42	91.89 ± 0.55	92.84 ± 0.29
	ECODv2	92.50 ± 0.32	93.61 ± 0.57	92.25 ± 0.50	92.91 ± 0.28
	IsolationForest	92.68 ± 0.31	93.87 ± 0.51	92.31 ± 0.47	93.07 ± 0.28
	OC-SVM	**92.71 ± 0.31**	93.76 ± 0.47	92.48 ± 0.46	**93.10 ± 0.27**
0.50	Baseline	86.37 ± 0.59	80.47 ± 0.42	**98.25 ± 0.65**	88.47 ± 0.49
	Baseline (orig)	88.74 ± 0.58	84.87 ± 0.74	96.01 ± 0.26	90.08 ± 0.47
	LBE	72.72 ± 0.43	92.31 ± 0.55	53.13 ± 0.74	67.41 ± 0.58
	SAR-EM	70.51 ± 0.24	**95.48 ± 0.31**	46.76 ± 0.25	62.77 ± 0.19
	A^3	92.75 ± 1.11	92.72 ± 1.32	93.91 ± 0.65	93.29 ± 0.95
	ECODv2	**92.93 ± 1.04**	93.21 ± 1.19	93.61 ± 0.91	93.39 ± 0.94
	IsolationForest	92.92 ± 1.02	93.07 ± 1.17	93.81 ± 0.74	**93.42 ± 0.89**
	OC-SVM	92.80 ± 1.05	93.03 ± 1.23	93.62 ± 0.66	93.31 ± 0.91
0.70	Baseline	90.20 ± 0.68	85.78 ± 0.74	**97.85 ± 0.55**	91.40 ± 0.57
	Baseline (orig)	90.55 ± 0.52	88.58 ± 0.71	94.46 ± 0.46	91.41 ± 0.44
	LBE	82.33 ± 0.50	88.17 ± 1.21	77.47 ± 1.21	82.33 ± 0.45
	SAR-EM	80.45 ± 0.21	94.92 ± 0.35	66.81 ± 0.23	78.42 ± 0.21
	A^3	93.54 ± 0.79	93.73 ± 0.83	94.22 ± 0.72	93.96 ± 0.71
	ECODv2	93.55 ± 0.73	93.86 ± 0.75	94.08 ± 0.88	93.95 ± 0.68
	IsolationForest	94.02 ± 0.62	**95.06 ± 0.46**	93.65 ± 0.89	94.33 ± 0.58
	OC-SVM	**94.05 ± 0.66**	94.81 ± 0.47	93.97 ± 0.87	**94.38 ± 0.62**

Table A4: Training time per dataset ($c = 0.5$).

Method	MNIST 3v5	MNIST OvE	CIFAR CarTruck	CIFAR MachineAnimal	STL MachineAnimal	Gas Concentrations
Baseline (modified)	258.77s	1617.03s	239.76s	1527.38s	255.23s	**74.90s**
Baseline (original)	**244.85s**	**1272.71s**	**215.28s**	**1037.12s**	**238.16s**	77.25s
SAR-EM	873.65s	44142.47s	9232.45s	47772.81s	5278.49s	87.37s
LBE	668.86s	12958.04s	927.65s	8984.34s	1090.33s	697.30s
VAE-PU+A^3	282.06s	1797.25s	256.86s	1635.81s	272.72s	87.80s
VAE-PU+ECOD	285.92s	1799.67s	256.44s	1645.48s	273.10s	82.40s
VAE-PU+IsolationForest	289.93s	1935.04s	264.96s	1673.16s	273.69s	86.90s
VAE-PU+OC-SVM	369.05s	3334.41s	249.64s	1798.46s	275.99s	79.30s

way leads to deterioration of the generative properties of the VAE-PU. Results of one of the tests conducted by us, illustrating the generation process on the MNIST 3v5 dataset, are shown in Table A1. Items were labeled based on the digit boldness, with bolder digits having a higher chance on being labeled. It is apparent that h_y-based sample matching leads to generation of the samples more similar (in terms of boldness) to the true PU set than h_s-based sample matching, where the generated items were more similar to PL set.

B Dataset details

For CIFAR-10 dataset, similarly to the original VAE-PU [16] approach, pretrained embedding vectors from Information Invariant Clustering (IIC; [11]) are used to extract features from images. Pretrained IIC embeddings were used also for STL-10 dataset. As MNIST and Gas Concentrations are significantly simpler, they are treated as tabular datasets, with only simple preprocessing (feature scaling). All of the benchark datasets are summarized in Table A2. Note that even though the classes are balanced in base data, labeling process (described later) naturally introduces various amounts of class imbalance, controlled by label frequency value c. This in particular results in a strong imbalance for small c; e.g. for CIFAR CarTruck data set and $c = 0.02$ one obtains, on average, 120 labeled examples and 11880 unlabeled ones.

C Precision-recall balance

Table A3 presents detailed experimental results on MNIST 3v5 dataset, which reveals interesting details concerning the precision and the recall for each of the tested methods. Note that in case of SAR-EM and LBE, the precision is really high and stays nearly constant, regardless of label frequency; on the other hand, for low c values the recall of the methods is extremely poor, indicating far too small number of positive predictions. Conversely, VAE-PU (in both versions) tends to skew the precision-recall balance the other way – even though the recall of the method is really remarkable, high number of False Positives means that both the precision and the overall predictive power of the model suffer. OCC-based sample selection helps VAE-PU achieve equilibrium of the precision and the recall. This effect is more noticeable for higher label frequencies – in all cases we can see a significant precision increase at the expense of slight recall decrease, but starting at $c = 0.10$ it causes precision and recall values to be nearly equal. Such behavior is consistent for all of the benchmark datasets, therefore the precision and the recall values were reported only in this case to avoid visual clutter in the tables. Relatively small precision values for low label frequencies might be explained by the imperfections of generated examples – when portion of labeled positives is small, it is harder to learn a robust positive data representation, as the labeled sample is poorly representative of the positive distribution due to biased PU problem nature. The generative process itself also emphasizes this issue – small labeled sample causes the generated examples to be limited in diversity, as discussed at the end of chapter 3.3.

D Training time comparison

Table A4 contains information about typical training times for a given algorithm on each dataset. Modified VAE-PU model tends to train slower than the original version, and to no surprise training time increases when using VAE-PU+OCC as opposed to the baseline; it should be emphasized, though, that the extra training step does not come at a heavy computational cost, as the typical training time increase ranges from 10% to 20%. A significant outlier is One-Class SVM variant, which can be slow for large datasets – this can be seen especially for MNIST OvE dataset, where the training time doubled after OC-SVM training. This example also highlights a significant advantage of VAE-PU-based models over the SAR-EM and LBE algorithm; when training dataset is not small, SAR-EM can reach training times up to 40-50 times larger than the other competitors, while LBE training is still comparably long even for small datasets. This property makes using VAE-PU and VAE-PU+OCC significantly more attractive even in high label frequency settings, where they offer way lower computational time combined with classification accuracy is on par or better than that of SAR-EM and LBE.

E Performance in SCAR setting

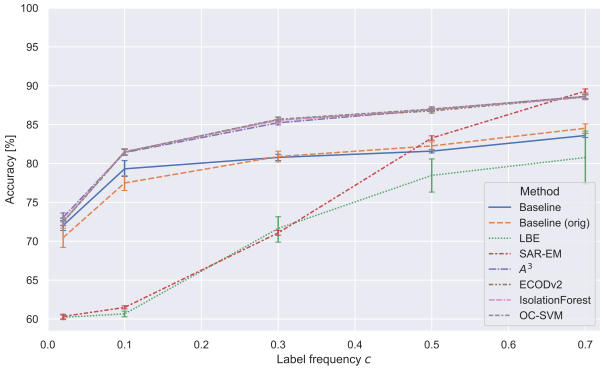

Figure A1: Test accuracy in SCAR setting, STL MachineAnimal dataset.

As the presented approach does not make any assumptions on the nature of biased sampling it is conjectured that it will also perform well under SCAR. In order to verify the hypothesis that OCC variants perform well even in SCAR setting, we performed additional tests

on STL dataset for this scenario. Here, items in the training set are labeled according to the SCAR assumption, i.e. probability of being labeled for each positive sample (propensity score) is constant and equal to label frequency c. As SCAR scenario is a special case of no-SCAR PU learning, it is expected that the results will stay similar to no-SCAR experiments.

Figure A1 illustrates results for the experiments in SCAR setting. The initial assumptions are confirmed – for all label frequencies, OCC variants outperform the baseline VAE-PU models, whereas SAR-EM performs poorly for low c values, while being competitive or even slightly outperforming the competition in high label frequency setting; LBE was outperformed in all test cases. Note that in many use cases the minor increase in classification performance of SAR-EM might be outweighed by a severe training time increase; nevertheless, some critical applications might find the high computational cost feasible. Also, performance of various OCC variants is almost indistinguishable in this case. Overall, the results follow similar patterns to no-SCAR scenario, which proves the effectiveness of VAE-PU+OCC even in a SCAR setting.

F T-test p-values

The tables A5 and A6 contain p-values of t-test of the null hypothesis that the accuracy or, respectively, F1 measure of the specific VAE-PU-OCC variant is equal to that of the baseline original method, against the alternative that it is larger than the baseline. Green ticks correspond to the cases when the null is rejected at $\alpha = 0.05$ significance level and dashes signify the failure to reject. As the averages of the metrics are based on 10 repetitions of the experiment, the benchmark distribution under the null hypothesis is t distribution with $10 + 10 - 2 = 18$ degrees of freedom.

G Properties of reverse sigmoid and logistic loss

G.1 Reverse sigmoid loss

We consider $\ell_{rsig}(x) = \left(1 + e^{-x}\right)^{-1}$. It is easy to check that it is monotone and convex. Consider the associated risk

$$R_{sig}(f) = E_X E_{Y|X=x}\, \ell_{rsig}(f(Yx)) := E_X R_{rsig}(f \mid X = x)$$

The risk given $X = x$ equals

$$\begin{aligned} R_{rsig}(f \mid X = x) = P(Y = 1) \left(1 + e^{-f(x)}\right)^{-1} \\ + P(Y = -1) \left(1 + e^{f(x)}\right)^{-1} \end{aligned}$$

Taking the derivative with respect to $f(x)$ and equating it to 0 we obtain

$$\begin{aligned} \frac{\partial R_{rsig}(f \mid X = x)}{\partial f(x)} = P(Y = 1 \mid x) \frac{e^{-f(x)}}{\left(1 + e^{-f(x)}\right)^2} \\ - P(Y = -1 \mid x) \frac{e^{f(x)}}{\left(1 + e^{f(x)}\right)^2}. \end{aligned}$$

Equating this to 0, yields the equality

$$\frac{P(Y = 1 \mid x)}{P(Y = -1 \mid x)} = e^{2f(x)} \left(\frac{1 + e^{-f(x)}}{1 + e^{f(x)}}\right)^2 \equiv 1.$$

Thus the stationary point does not exist if $P(Y = 1 \mid x) \not\equiv \frac{1}{2}$ and if the condition holds any $f(x)$ is a stationary point (which is obvious as risk function is equal to 1 in this case).

G.2 Logistic loss

We consider $\ell_{logist}(x) = \log\left(1 + e^{-x}\right)$. It is easy to check that it is monotone and convex. We obtain

$$\begin{aligned} R_{logist}(f \mid X = x) = P(Y = 1 \mid X = x) \log\left(1 + e^{-f(x)}\right) \\ + P(Y = -1 \mid X = x) \log\left(1 + e^{f(x)}\right) \end{aligned}$$

and stationary point f^* satisfies

$$\frac{P(Y = 1 \mid X = x)}{P(Y = -1 \mid X = x)} = e^{2f(x)} \frac{1 + e^{-f(x)}}{1 + e^{f(x)}} = e^{f(x)}$$

and thus

$$f^*(x) = \log \frac{P(Y = 1 \mid X = x)}{P(Y = -1 \mid X = x)}$$

and as monotone function of posterior odds yields a Bayes rule.

Table A5: T-test p-values for Accuracy per dataset.

c	Method	MNIST 3v5	MNIST OvE	CIFAR CarTruck	CIFAR MachineAnimal	STL MachineAnimal	Gas Concentrations
0.02	A^3	0.09 –	< 0.01 ✓	0.16 –	0.03 ✓	< 0.01 ✓	< 0.01 ✓
	IsolationForest	0.06 –	< 0.01 ✓	0.01 ✓	0.07 –	< 0.01 ✓	0.03 ✓
	ECODv2	0.07 –	< 0.01 ✓	0.41 –	0.04 ✓	0.02 ✓	0.03 ✓
	OC-SVM	0.06 –	< 0.01 ✓	0.38 –	0.04 ✓	0.02 ✓	0.02 ✓
0.10	A^3	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	0.06 –	< 0.01 ✓
	IsolationForest	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	0.07 –	0.18 –
	ECODv2	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	0.09 –	0.02 ✓
	OC-SVM	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	0.08 –	0.07 –
0.30	A^3	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓
	IsolationForest	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓
	ECODv2	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓
	OC-SVM	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓
0.50	A^3	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓
	IsolationForest	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	0.01 ✓
	ECODv2	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓
	OC-SVM	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	0.01 ✓
0.70	A^3	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓
	IsolationForest	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓
	ECODv2	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓
	OC-SVM	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓

Table A6: T-test p-values for F1 score per dataset.

c	Method	MNIST 3v5	MNIST OvE	CIFAR CarTruck	CIFAR MachineAnimal	STL MachineAnimal	Gas Concentrations
0.02	A^3	0.09 –	< 0.01 ✓	0.09 –	0.03 ✓	< 0.01 ✓	< 0.01 ✓
	IsolationForest	0.11 –	< 0.01 ✓	0.01 ✓	0.08 –	0.05 ✓	0.09 –
	ECODv2	0.10 –	< 0.01 ✓	0.29 –	0.05 ✓	0.08 –	0.07 –
	OC-SVM	0.11 –	< 0.01 ✓	0.28 –	0.05 ✓	0.08 –	0.03 ✓
0.10	A^3	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	0.04 ✓	< 0.01 ✓
	IsolationForest	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	0.04 ✓	0.29 –
	ECODv2	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	0.06 –	0.03 ✓
	OC-SVM	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	0.05 –	0.16 –
0.30	A^3	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓
	IsolationForest	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓
	ECODv2	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓
	OC-SVM	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓
0.50	A^3	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓
	IsolationForest	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	0.01 ✓
	ECODv2	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓
	OC-SVM	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	0.01 ✓
0.70	A^3	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓
	IsolationForest	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓
	ECODv2	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓
	OC-SVM	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓	< 0.01 ✓