Augmented prediction of a true class for Positive Unlabeled data under selection bias – supplementary materials

Jan Mielniczuk[a,b,*,1] **and Adam Wawrzeńczyk**[a,1]

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

A Example 1 – full derivation

Example 1. *Let $y(x) = \Phi(x), X \sim N(0,1)$, and $x \in \mathbb{R}$ (univariate probit model with standard normal predictor), and let propensity score $e_a(x) = \mathbb{I}\{x > a\}$ i.e. above threshold $a \in \mathbb{R}$ all positive observations are labeled. It is easy to check that*

$$P(Y=1) = \int P(Y=1|X=x)f(x)\ \mathrm{d}x = \int_{-\infty}^{\infty} \Phi(x)\phi(x)\ \mathrm{d}x = \int_0^1 z\ \mathrm{d}z = \frac{1}{2}.$$

and Bayes risk of $d_B(x)$ equals

$$L^* = \int_{-\infty}^{\infty} \min\big(\Phi(x), 1-\Phi(x)\big)\phi(x)\ \mathrm{d}x = \int_{-\infty}^{0} \Phi(x)\phi(x)\ \mathrm{d}x + \int_0^{\infty} (1-\Phi(x))\phi(x)\ \mathrm{d}x = \frac{1}{4}.$$

As $s(x) = y(x)\mathbb{I}\{x > a\}$, probability of labeling equals

$$P(S=1) = \int_{-\infty}^{\infty} P(S=1|X=x)f(x)\ \mathrm{d}x = \int_{-\infty}^{\infty} s(x)\phi(x)\ \mathrm{d}x = \int_a^{\infty} \Phi(x)\phi(x)\ \mathrm{d}x = \frac{1}{2}(1-\Phi^2(a))$$

and $P(S=0) = \frac{1}{2}(1+\Phi^2(a))$. Moreover, $L^*_{PU} = L^*_{PU}(a)$ equals for $a > 0$

$$L^*_{PU}(a) = \mathbb{E}_{X,S=0} \min\big(\tilde{y}(X,0), 1-\tilde{y}(X,0)\big) = \left\{\int_{-\infty}^{0} \Phi(x)\phi(x)\ \mathrm{d}x + \int_0^{a} (1-\Phi(x))\phi(x)\ \mathrm{d}x\right\} = \Phi(a) - \frac{\Phi^2(a)}{2} - \frac{1}{4}.$$

and analogous calculation for $a \leq 0$ yields $L^*_{PU}(a) = \Phi^2(a)/2$.

Thus the excess risk of $d_B(x)$ defined in (3) for $a > 0$ equals

$$\mathbb{E}_X\left[\min\big(y(X), 1-y(X)\big)\right] - \mathbb{E}_{X,S}\left[\min\big(\tilde{y}(X,S), 1-\tilde{y}(X,S)\big)\right] = \frac{1}{2} - \Phi(a) + \frac{\Phi^2(a)}{2} = \frac{1}{2}\big(\Phi(a)-1\big)^2 \geq 0,$$

and for $a < 0$ equals $\frac{1}{4} - \frac{\Phi^2(a)}{2} \geq 0$. Note that for $a \to \infty$ excess risk tends to 0 as $P_{X,S=0}$ approaches P_X in this case and $d_B^{PU}(x,0)$ tends to $d_B(x)$. For $a \to -\infty$ the excess risk tends to 1/4 (risk of $d_B(x)$) as the risk of $d_B^{PU}(x,s)$ tends to 0.

B Experiment dataset statistics

Table A1 summarized the properties of all datasets used in the experiment section of the paper.

Table A1. 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
CDC Diabetes	148458	38	0.50

C Balanced accuracy results

Tables A2 through A5 correspond to tables 2-5 in the main paper and present U-balanced accuracy for all the performed experiments. This is an important metric, as U-metrics introduce imbalance into the measurements, which might impact accuracy negatively. We can see, however, that the obtained results are close to results presented in the main paper – while VP-B+S does not have as overwhelming of an advantage, it still clearly is the best method overall.

* Corresponding Author. Email: jan.mielniczuk@ipipan.waw.pl.
[1] Equal contribution.

Table A2. U-Balanced accuracy values – Method comparison – Synthetic datasets

c	Method	Synth. 1	Synth. 2	Synth. 3	Synth. SCAR
0.02	VP+S	**61.29 ± 2.27**	**59.43 ± 2.66**	**60.53 ± 2.30**	**63.52 ± 1.77**
	VP-B+S	61.04 ± 2.33	59.32 ± 2.61	60.40 ± 2.25	63.33 ± 1.73
	LBE+S	49.70 ± 0.35	50.01 ± 0.35	49.93 ± 0.32	49.69 ± 0.33
0.10	VP+S	67.67 ± 0.52	**66.72 ± 0.60**	**67.47 ± 0.61**	**68.49 ± 0.47**
	VP-B+S	**67.71 ± 0.52**	66.45 ± 0.54	67.43 ± 0.58	68.26 ± 0.47
	LBE+S	50.03 ± 0.33	50.83 ± 0.36	50.22 ± 0.32	49.66 ± 0.37
0.30	VP+S	67.49 ± 0.58	65.39 ± 0.59	66.24 ± 0.50	**70.33 ± 0.56**
	VP-B+S	**67.74 ± 0.56**	**65.61 ± 0.63**	**66.35 ± 0.52**	70.17 ± 0.55
	LBE+S	52.86 ± 0.18	52.88 ± 0.33	52.68 ± 0.22	50.00 ± 0.32
0.50	VP+S	**64.11 ± 0.54**	**61.68 ± 0.62**	63.18 ± 0.67	**69.30 ± 0.43**
	VP-B+S	64.09 ± 0.53	61.65 ± 0.61	**63.27 ± 0.70**	69.21 ± 0.44
	LBE+S	55.57 ± 0.49	54.62 ± 0.47	54.96 ± 0.50	56.69 ± 0.44
0.70	VP+S	**59.40 ± 0.78**	58.33 ± 0.73	**58.51 ± 0.65**	65.79 ± 0.56
	VP-B+S	**59.40 ± 0.79**	**58.42 ± 0.69**	58.35 ± 0.60	65.49 ± 0.71
	LBE+S	58.33 ± 0.65	56.61 ± 0.77	56.60 ± 0.66	**67.87 ± 0.59**
0.90	VP+S	52.19 ± 0.57	52.48 ± 0.58	52.33 ± 0.51	59.52 ± 0.98
	VP-B+S	52.19 ± 0.55	52.41 ± 0.57	52.48 ± 0.60	58.64 ± 0.85
	LBE+S	**57.42 ± 0.86**	**56.47 ± 0.92**	**54.63 ± 1.01**	**71.37 ± 1.02**

Table A3. U-Balanced accuracy values – Method comparison – Real-world datasets

c	Method	MNIST 3v5	MNIST OvE	CIFAR CT	CIFAR MA	STL MA	CDC-Diabetes
0.02	VP+S	76.89 ± 1.14	68.32 ± 1.19	87.28 ± 0.49	91.24 ± 0.21	82.13 ± 0.64	50.00 ± 1.48
	VP-B+S	**78.37 ± 1.53**	**74.88 ± 1.51**	**92.46 ± 0.43**	**94.03 ± 0.09**	**83.42 ± 0.69**	**51.16 ± 1.74**
	LBE+S	49.95 ± 0.34	50.14 ± 0.14	50.00 ± 0.40	50.16 ± 0.18	50.17 ± 0.24	49.97 ± 0.20
0.10	VP+S	80.06 ± 0.61	74.69 ± 1.38	91.64 ± 0.32	92.75 ± 0.24	87.19 ± 0.32	57.91 ± 0.78
	VP-B+S	**84.29 ± 0.76**	**83.22 ± 1.23**	**93.46 ± 0.19**	**94.28 ± 0.13**	**88.24 ± 0.30**	**60.93 ± 0.68**
	LBE+S	50.25 ± 0.34	49.79 ± 0.15	50.61 ± 0.38	50.34 ± 0.28	50.78 ± 0.28	49.93 ± 0.19
0.30	VP+S	82.31 ± 0.67	80.74 ± 0.86	93.20 ± 0.26	94.06 ± 0.10	**88.65 ± 0.28**	57.47 ± 0.84
	VP-B+S	**86.95 ± 0.51**	**87.99 ± 0.64**	**94.20 ± 0.16**	**94.78 ± 0.09**	88.32 ± 0.35	**61.60 ± 0.92**
	LBE+S	50.77 ± 0.31	49.79 ± 0.18	55.32 ± 1.30	60.51 ± 4.24	55.90 ± 1.85	49.93 ± 0.17
0.50	VP+S	85.05 ± 0.70	84.64 ± 0.78	94.24 ± 0.24	**94.52 ± 0.12**	**89.34 ± 0.47**	59.37 ± 0.56
	VP-B+S	**88.49 ± 0.83**	**89.86 ± 0.59**	**94.73 ± 0.24**	94.43 ± 0.15	88.60 ± 0.53	**62.98 ± 0.41**
	LBE+S	51.80 ± 0.34	52.90 ± 1.30	72.38 ± 2.72	73.44 ± 2.77	70.34 ± 2.07	62.39 ± 1.83
0.70	VP+S	88.28 ± 0.67	89.25 ± 0.54	94.35 ± 0.32	**94.23 ± 0.17**	**88.63 ± 0.54**	60.07 ± 0.73
	VP-B+S	**90.09 ± 0.64**	**91.30 ± 0.55**	**94.66 ± 0.27**	93.51 ± 0.26	87.97 ± 0.51	62.31 ± 0.62
	LBE+S	54.66 ± 0.65	60.61 ± 1.00	91.25 ± 1.47	90.31 ± 1.01	83.02 ± 1.48	**69.67 ± 1.02**
0.90	VP+S	**87.14 ± 0.74**	**92.30 ± 0.40**	91.97 ± 0.55	92.77 ± 0.45	82.75 ± 1.47	63.01 ± 0.71
	VP-B+S	86.15 ± 0.70	91.84 ± 0.60	92.00 ± 0.55	89.65 ± 0.34	82.43 ± 1.45	58.52 ± 0.53
	LBE+S	67.77 ± 0.98	83.49 ± 0.79	**93.50 ± 0.46**	**93.68 ± 0.46**	**88.35 ± 0.70**	**72.88 ± 0.38**

Table A4. U-Balanced accuracy values – Decision rule comparison – Synthetic datasets

c	Method	Synth. 1	Synth. 2	Synth. 3	Synth. SCAR
0.02	S-Prophet	73.29 ± 0.35	73.25 ± 0.36	71.37 ± 0.35	73.48 ± 0.35
	Y-Prophet	73.32 ± 0.36	73.25 ± 0.36	71.40 ± 0.36	73.51 ± 0.35
	VP-B	60.94 ± 2.39	**59.42 ± 2.55**	60.14 ± 2.31	**63.33 ± 1.68**
	VP-B+S	**61.04 ± 2.33**	59.32 ± 2.61	**60.40 ± 2.25**	**63.33 ± 1.73**
	VP-B+S + true s(x)	60.81 ± 2.36	59.16 ± 2.68	60.18 ± 2.29	63.23 ± 1.70
	VP-B+S + true y(x)	73.29 ± 0.37	73.22 ± 0.36	71.43 ± 0.36	73.46 ± 0.34
0.10	S-Prophet	72.52 ± 0.31	72.06 ± 0.35	70.49 ± 0.31	73.61 ± 0.35
	Y-Prophet	72.60 ± 0.35	72.13 ± 0.38	70.65 ± 0.37	73.69 ± 0.33
	VP-B	**68.02 ± 0.46**	**66.65 ± 0.57**	**67.62 ± 0.59**	68.16 ± 0.38
	VP-B+S	67.71 ± 0.52	66.45 ± 0.54	67.43 ± 0.58	**68.26 ± 0.47**
	VP-B+S + true s(x)	67.80 ± 0.52	66.57 ± 0.49	67.30 ± 0.62	68.20 ± 0.46
	VP-B+S + true y(x)	72.53 ± 0.35	71.75 ± 0.41	70.42 ± 0.36	73.55 ± 0.33
0.30	S-Prophet	69.91 ± 0.47	69.02 ± 0.52	67.65 ± 0.42	72.69 ± 0.48
	Y-Prophet	70.57 ± 0.43	69.37 ± 0.42	68.49 ± 0.39	73.60 ± 0.35
	VP-B	**68.08 ± 0.52**	**65.85 ± 0.58**	**67.19 ± 0.52**	**70.53 ± 0.49**
	VP-B+S	67.74 ± 0.56	65.61 ± 0.63	66.35 ± 0.52	70.17 ± 0.55
	VP-B+S + true s(x)	67.67 ± 0.57	65.70 ± 0.64	66.48 ± 0.55	70.14 ± 0.51
	VP-B+S + true y(x)	69.30 ± 0.47	68.33 ± 0.52	67.14 ± 0.43	72.61 ± 0.50
0.50	S-Prophet	65.33 ± 0.70	63.89 ± 0.56	63.25 ± 0.71	71.13 ± 0.62
	Y-Prophet	68.13 ± 0.49	66.56 ± 0.48	66.02 ± 0.50	73.18 ± 0.36
	VP-B	**64.95 ± 0.45**	**62.65 ± 0.56**	**64.12 ± 0.70**	**70.31 ± 0.49**
	VP-B+S	64.09 ± 0.53	61.65 ± 0.61	63.27 ± 0.70	69.21 ± 0.44
	VP-B+S + true s(x)	64.08 ± 0.48	61.58 ± 0.60	63.15 ± 0.67	69.30 ± 0.46
	VP-B+S + true y(x)	64.44 ± 0.62	62.68 ± 0.60	62.03 ± 0.82	70.47 ± 0.69
0.70	S-Prophet	57.88 ± 0.51	56.62 ± 0.38	56.71 ± 0.61	66.39 ± 0.48
	Y-Prophet	64.64 ± 0.70	63.34 ± 0.70	62.71 ± 0.73	73.39 ± 0.41
	VP-B	**61.12 ± 0.65**	**59.53 ± 0.68**	**59.84 ± 0.68**	**68.95 ± 0.68**
	VP-B+S	59.40 ± 0.79	58.42 ± 0.69	58.35 ± 0.60	65.49 ± 0.71
	VP-B+S + true s(x)	59.33 ± 0.61	58.45 ± 0.61	58.31 ± 0.70	65.01 ± 0.78
	VP-B+S + true y(x)	57.07 ± 0.73	56.58 ± 0.52	56.44 ± 0.64	65.76 ± 0.39
0.90	S-Prophet	50.38 ± 0.19	50.35 ± 0.21	50.47 ± 0.16	55.62 ± 0.55
	Y-Prophet	61.12 ± 1.16	60.29 ± 1.07	58.82 ± 1.08	72.72 ± 1.00
	VP-B	**58.85 ± 0.62**	**57.12 ± 0.69**	**55.19 ± 0.99**	**71.61 ± 0.91**
	VP-B+S	52.19 ± 0.55	52.41 ± 0.57	52.48 ± 0.60	58.64 ± 0.85
	VP-B+S + true s(x)	52.99 ± 0.40	52.90 ± 0.41	51.86 ± 0.67	60.43 ± 0.93
	VP-B+S + true y(x)	50.94 ± 0.53	51.03 ± 0.41	52.25 ± 0.50	54.80 ± 0.55

Table A5. U-Balanced accuracy values – Decision rule comparison – Real-world datasets

c	Method	MNIST 3v5	MNIST OvE	CIFAR CT	CIFAR MA	STL MA	CDC-Diabetes
0.02	VP-B	78.32 ± 1.45	74.49 ± 1.49	92.42 ± 0.41	93.91 ± 0.09	**83.46 ± 0.68**	51.12 ± 1.76
	VP-B+S	**78.37 ± 1.53**	**74.88 ± 1.51**	**92.46 ± 0.43**	**94.03 ± 0.09**	83.42 ± 0.69	**51.16 ± 1.74**
0.10	VP-B	84.10 ± 0.65	82.80 ± 1.28	93.35 ± 0.19	**94.28 ± 0.12**	**88.27 ± 0.29**	**61.38 ± 0.77**
	VP-B+S	**84.29 ± 0.76**	**83.22 ± 1.23**	**93.46 ± 0.19**	**94.28 ± 0.13**	88.24 ± 0.30	60.93 ± 0.68
0.30	VP-B	86.75 ± 0.55	**88.04 ± 0.65**	94.12 ± 0.18	**94.35 ± 0.08**	**89.11 ± 0.28**	**63.78 ± 0.78**
	VP-B+S	**86.95 ± 0.51**	87.99 ± 0.64	**94.20 ± 0.16**	94.78 ± 0.09	88.32 ± 0.35	61.60 ± 0.92
0.50	VP-B	88.33 ± 0.88	**90.09 ± 0.57**	94.69 ± 0.25	**94.72 ± 0.10**	**89.81 ± 0.34**	**67.36 ± 0.25**
	VP-B+S	**88.49 ± 0.83**	89.86 ± 0.59	**94.73 ± 0.24**	94.43 ± 0.15	88.60 ± 0.53	62.98 ± 0.41
0.70	VP-B	**90.23 ± 0.66**	**92.15 ± 0.47**	**94.81 ± 0.26**	**94.65 ± 0.15**	**90.82 ± 0.36**	**68.93 ± 0.40**
	VP-B+S	90.09 ± 0.64	91.30 ± 0.55	94.66 ± 0.27	93.51 ± 0.26	87.97 ± 0.51	62.31 ± 0.62
0.90	VP-B	**87.31 ± 0.72**	**93.55 ± 0.45**	**94.07 ± 0.33**	**94.89 ± 0.19**	**90.03 ± 0.73**	**71.38 ± 0.37**
	VP-B+S	86.15 ± 0.70	91.84 ± 0.60	92.00 ± 0.55	89.65 ± 0.34	82.43 ± 1.45	58.52 ± 0.53